N° d'ordre : 262/20

UNIVERSITÉ SULTAN MOULAY SLIMANE
FACULTÉ DES SCIENCES ET TECHNIQUES
Béni Mellal

Centre des Etudes Doctorales : Sciences et Techniques.
Formation doctorale : Mathématique et Physique appliquées.

THÈSE

Presentée par

Mohamed EL MOHADAB
(Master : Ingénierie Informatique et Systémes)

Pour l'obtention du grade de

DOCTORAT

Specialité : Informatique.

Nouvelle Approche des Algorithmes de Prédiction et de Classification : Application à la Recherche Scientifique Universitaire

Soutenue le 25/07/2020 devant le jury composé de :

Pr. Mostafa JOURHMANE	: Professeur Faculté des Sciences et Techniques de Béni Mellal	Président du jury.
Pr. Benayad NSIRI	: Professeur Ecole Normale Supérieure de l'Enseignement Technique de Rabat	Rapporteur.
Pr. Hicham MOUNCIF	: Professeur Faculté Polydisciplinaire de Béni Mellal	Rapporteur.
Pr. Mohamed BASLAM	: Professeur Faculté des Sciences et Techniques de Béni Mellal	Rapporteur.
Pr. Said SAFI	: Professeur Faculté Polydisciplinaire de Béni Mellal	Directeur de la thèse.
Pr. Belaid BOUIKHALENE	: Professeur Faculté Polydisciplinaire de Béni Mellal	Co-directeur de thèse.

Remerciements

J'aimerais tout d'abord remercier mon directeur de thèse le professeur monsieur **Said SAFI** pour ses conseils et surtout son sens pédagogique m'ont permis de trouver la force de mener à bout ce long projet. Ainsi pour la confiance qu'il m'a accordée en acceptant d'encadrer ce travail doctoral, pour ses multiples conseils et pour tout le temps qu'il a consacré à diriger cette recherche. J'aimerais également lui dire à quel point j'ai apprécié sa grande disponibilité et son respect sans faille des délais serrés de relecture des documents que je lui ai adresses.

J'aimerais également remercier profondément mon co-directeur, monsieur **Belaid BOUIKHALENE** qui m'a inculqué les principes de la recherche et qu'il m'a patiemment amené à formaliser les idées qui sont au cœur de ce travail. Au cours de nos nombreux entretiens, j'ai apprécié son écoute, sa rigueur et la profondeur de ses connaissances. Aussi, je le remercie sincèrement pour son encadrement considérable, ses idées précieuses et son soutien constant ainsi que son assistance morale qui m'a été d'une utilité inestimable.

Merci surtout à ma mère et mon père, qui m'ont toujours épaulé dans mes études et qui m'ont toujours encouragé à atteindre les objectifs que je me fixais. Le plus grand des mercis à mes parents, pour ses indéfectible support durant cette période et pour le soutien qu'ils m'ont manifesté chaque jour de ma vie. Ainsi je remercie ma sœur et mon frère qui m'ont apporté leur aide nécessaire au cours de mes études et qui m'ont supporté dans les moments de stress et de difficultés.

Je remercie vivement tous les membres du jury de ma thèse qui ont pris de leur temps pour lire et juger mon travail ainsi que pour leur déplacement le jour de la soutenance. Je remercie M. **Mostafa JOURHMANE** pour l'honneur qu'il m'a fait en acceptant de présider le jury de ma thèse. Je remercie, également, infiniment les rapporteurs de ma thèse, M. **Benayad NSIRI** professeur à l'Ecole Normale Superieure de l'Enseignement Technique de Rabat , M. **Hicham MOUNCIF** professeur à la Faculté Polydisciplinaire de Béni Mellal et M. **Mohamed BASLAM** professeur à la Faculté des Sciences et Techniques de Béni Mellal, pour avoir consacré du temps à la lecture de cette thèse ainsi pour avoir soumis leur précieux jugement sur la qualité et le contenu de ce travail.

Enfin, je remercie toutes les personnes qui, de près ou de loin, ont apporté leur contribution à ce travail. Je leur exprime ici toute ma reconnaissance et ma sympathie.

Avant-propos

0.1 Renseignements de la thèse

-Prénom et Nom de l'auteur de la thèse : Mohamed EL MOHADAB.

-Titre de la thèse : Nouvelle Approche des Algorithmes de Prédiction et de Classification : Application à la Recherche Scientifique Universitaire.

-Prénom et Nom du directeur de la thèse : Pr. Said SAFI (Professeur à la Faculté Polydisciplinaire de Béni-Mellal).

-Laboratoire d'acceuil : Laboratoire d'Innovation en Mathématiques, Applications & Technologies de l'Information (LIMATI) , Faculté Polydisciplinaire, Université Sultan Moulay Slimane, Béni-Mellal.

0.2 Liste des Publications

0.2.1 Articles dans des journaux internationaux

1. **Mohamed El Mohadab**, Belaid Bouikhalene, and Said Safi. Predicting rank for scientific research papers using supervised learning. Applied Computing and Informatics, Elsevier, 15(2) : 182-190, 2018.

2. **Mohamed El Mohadab**, Belaid Bouikhalene, and Said Safi. Automatic CV processing for scientific research using data mining algorithm. Journal of King Saud University - Computer and Information Sciences, Elsevier, 32(5) : 561-567, 2020.

3. **Mohamed El Mohadab** , Belaid Bouikhalene, Fahd Ouatik and Said Safi. AHP and TOPSIS methods applied in the field of scientific research. Indonesian Journal of Electrical Engineering and Computer Science, 14(3) : 1382-1390, 2019.

4. **Mohamed El Mohadab** , Belaid Bouikhalene, and Said Safi. Bibliometric method for mapping the state of the art of scientific production in Covid-19. Chaos, Solitons & Fractals, 139 : 110052, 2019.

0.2.2 Proceedings de conférences internationales

5. **M. El Mohadab**, Belaid Bouikhalene, and Said Safi. Decision making system for scientific research using data mining algorithm. In International Arab Conference on Information Technology (ACIT), December 2016.

6. **M. El Mohadab**, Belaid Bouikhalene, and Said Safi. Toward an efficient algorithm for ranking scientific research papers. In Scientific Event on Information Technology (SEIT), 2017, May 2017.
7. **M. El Mohadab**, Belaid Bouikhalene, and Said Safi. Enterprise resource planning : introductory overview. In International Conference on Electrical and Information Technologies (ICEIT), IEEE, November 2017.
8. **M. El Mohadab**, Belaid Bouikhalene, and Said Safi. Impact of enterprise resource planning systems on scientific research system in public university. In International Arab Conference on Information Technology (ACIT), December 2017.

9. **M. El Mohadab**, Belaid Bouikhalene, and Said Safi. Adaptive ranking of scientific research papers based on author score and paper score. In International Conference on signals, Automatic and Telecommunications (ICSAT), May 2018.

10. **M. El Mohadab**, Belaid Bouikhalene, and Said Safi. Integration between enterprise resource planning systems and data warehousing. In International Conference on signals, Automatic and Telecommunications (ICSAT), May 2018.

11. **M. El Mohadab**, Participation au forum du 10ème Edition du Prix National de l'Administration Electronique e-mtiaz 2016 à Rabat par un projet intitulé Management de la recherche à l'Université Sultan Moulay Slimane organisée le 21 Décembre 2016 à Rabat.

Résumé

Cette thèse porte sur de nouvelles approches pour faire évoluer un système d'information universitaire en un système d'information décisionnelle universitaire, afin de développer les processus de la recherche scientifique dans une université publique. Le transfert d'un système d'information en système d'information décisionnel repose sur les bases métiers orientées vers les chercheurs de l'université par la prise en compte de la modélisation des intervenants, l'enjeu est d'étudier, de modéliser et d'automatiser le processus afin de permettre au responsable universitaire de prendre la meilleure décision stratégiques et instantanée. Par notre contribution, nous étudions les processus propres de la recherche scientifique en intégrant les outils d'exploration de données et la recherche opérationnelle.

Les avancées rapides des technologies de l'information et de la communication ont des conséquences capitales sur l'évolution de la recherche scientifique. À ce propos, l'évolution du système d'information universitaire recouvre tous les processus relatifs au système d'information de la recherche scientifique d'une université publique qui s'appuie sur les technologies de l'information et de la communication. L'enjeu est de fournir un contenu adapté aux attentes des chercheurs et les dirigeants. La majorité des systèmes de la recherche scientifique d'aujourd'hui manque de méthodes pour assister les besoins des chercheurs qui sont généralement hétérogènes en termes de diversité, préférences, etc. D'autre part, pour répondre aux exigences des dirigeants il faut alors fournir des mécanismes puissants et adéquat pour adapter les décisions (pédagogiques, financiers, administratives, ...) aux besoins particuliers de chaque établissement, et ces structures de recherche.

Notre contribution dans ce domaine de recherche porte sur le développement d'une plateforme de recherche scientifique qui permet le management du processus propre à la recherche scientifique. Nous avons étudié la problématique de l'adaptation comme un problème d'optimisation, en utilisant les algorithmes de l'exploration de données qui sont fondés sur la théorie de prédiction.

En outre, nous proposons une méthode de prédiction du classement du papier scientifique, considéré comme un sous-ensemble de la gestion de la recherche scientifique. Cette méthode de classement permet de prédire le classement du papier en se basant sur le classement des anciens papiers scientifiques publiés et susceptibles d'intéresser le chercheur. Une telle recommandation est basée sur une méthode d'apprentissage supervisée qui combine entre des données relatives au papier scientifique et l'apprentissage automatique. Notre objectif principal est d'orienter les chercheurs et leurs suggérer le

futur classement de leur papier à la base de leurs expériences d'apprentissage.
Par ailleurs, la gestion intelligente des curriculums vitæ des chercheurs est une source de données pour les concepteurs. C'est la raison pour laquelle nous avons conçu une solution qui vise essentiellement à prédire le domaine de recherche de chaque chercheur à travers les données tirées du curriculum vitæ. Cette méthode est fondée sur le mixage entre le traitement du langage naturel et l'apprentissage supervisé. A travers l'application des algorithmes qui relèvent du domaine de l'apprentissage automatique pour construire un modèle prédictif basé sur les arbres de décision.
Ainsi, la prise de décision par un décideur nécessite l'appuie sur des bonnes pratiques qui finalisent un processus qui mène à la bonne gouvernance et la gestion intelligente, spécialement dans le côté financier. Pour bénéficier des points forts de cette direction, nous avons appliqué des solutions qui relèvent des méthodes de prise de décision à critères multiples qui sont connues par leurs fiabilité et crédibilités dans ce genre d'étude.

Mots-clés : Recherche scientifique, Progiciel de gestion intégré, Classement du papier scientifique, Traitement automatique de la langue, Exploration de données, Apprentissage supervisé, Prédiction, Méthodes de prise de décision à critères multiples.

Abstract

This thesis deals with new approaches to evolve a university information system into an academic decision-making information system, in order to develop the processes of brick scientific research in a public university. The transfer of an information system into a decision-making information system is based on the university-oriented business bases of the university by taking into account stakeholder modeling, the issue and studying, modeling and automating the process to allow the academic lead to make the best strategic and instant decision. Through our contribution, we study the specific processes of scientific research by integrating data mining and operational research tools.

Rapid advances in information and communication technologies has a major impact on the evolution of scientific research. In this regard, the evolution of the university information system covers all the processes relating to the scientific research information system of a public university that rely on information and communication technologies. The challenge is to provide content adapted to the expectations of researchers and leaders. The majority of today's scientific research systems lack methods to support the needs of researchers who are generally heterogeneous in terms of diversity, preferences, and so on. On the other hand, it must respond to the demands of the leaders. It is then necessary to provide powerful and adequate mechanisms to adapt the pedagogical, financial, administrative, and other decisions to the particular needs of each institution and its research structures.

Our contribution in this area is the development of a scientific research platform that allows the management of the scientific research process itself. We have studied the problem of adaptation as an optimization problem, using data mining algorithms based on prediction theory. In addition, we propose a method for predicting the ranking of scientific paper, considered as a subset of the management of scientific research. This ranking method predicts the ranking of paper based on the ranking of published scientific papers that may be of interest to the researcher. Such a recommendation is based on a supervised learning method that combines between scientific paper data and machine learning. Our main goal is to guide researchers and suggest the future classification of their paper base on their learning experiences.

In addition, the intelligent management of researchers is a source of data for designers. This is why we have devised a solution that aims essentially to predict the field of work of each researcher through the data extracted from the curriculum vitae. This method is based on the mix between natural language processing and supervised learning. Through the application of algorithms that fall within the field of machine learning to build a pre-

dictive model based on decision trees. Thus, decision-making by a decision maker requires the support of good practices that finalizes a process that leads to good governance and smart management, especially in the financial side, to benefit from the strengths of this direction we have applied solutions that fall under multi-criteria decision-making methods, which are known by its reliability and credibility in this kind of study.

Keywords : Scientific research, Enterprise resource planning, Scientific paper ranking, Automatic language processing, Data mining, Supervised learning, Prediction, Multi-criteria decision making methods.

Table des matières

Table des figures

Liste des tableaux

Liste des Acronymes

THEWUR	Times Higher Education World University Rankings.
SCI	Science Citation Index.
HEFC	Higher Education Funding Councils.
ANVUR	Italian National Agency for the Evaluation of the University and Research Systems.
ANECA	National Agency for Quality Assessment and Accreditation.
AEQES	Agence pour l'Évaluation de la Qualité de l'Enseignement Supérieur..
AERES	Agence d'Évaluation de la Recherche et de l'Enseignement Supérieur.
CNCPRST	Centre National de Coordination et de Planification de la Recherche Scientifique et Technique .
CNRST	Centre National pour la Recherche Scientifique et Technique.
ANEAQ	Agence Nationale d'Évaluation et d'Assurance Qualité de L'Enseignement Supérieur et de la Recherche Scientifique .
ANQAHE	Arab Network for Quality Assurance in Higher Education.
FrAQ_Sup	Réseau Francophone des Agences Qualité pour l'Enseignement Supérieur .
IMIST	Institut Marocain de l'Information Scientifique et Technique.
WOS	Web Of Science.
SJR	Scimago Journal Rank.
SNIP	Source Normalized Impact per Paper.
TIC	Technologies de l'Information et de la Communication.
MRP	Materials Requirements Planning.
SI	Système d'Information.
PME	Petites et Moyennes Entreprises.
UML	Unified Modeling Language.
CV	Curriculum Vitae.
TALN	Traitement Automatique du Langage Naturel.
TF	Term Frequency.
Tf-Idf	Term Frequency-Inverse Document Frequency.
CCI	Correctly Classified Instances.
ICI	Incorrectly Classified Instances..
IR	Information Retrieval.
ERP	Enterprise Resource Planning.
PGI	Progiciel de Gestion Intégré.
HITS	Hyperlink-Induced Topic Search.
DOI	Digital Object Identifier.
nDCG	Normalized Discounted Cumulative Gain.
CG	Cumulative Gain.
MAP	Mean Average Precision.

GMAP	Geometric Mean Average Precision.
AHP	Analytic Hierarchy Process.
TOPSIS	Technique for Order of Preference by Similarity to Ideal Solution.
MCDM	Multiple Criteria Decision Making.
MODM	Multi Objective Decision Making.
MADM	Multi Attribute Decision Making.
CR	Consistency Ratio.
CI	Concistency Index.

Introduction générale

Contexte et motivation

Ce travail présente la synthèse de quatre années d'investigation effectuée dans le cadre d'une thèse de doctorat en Informatique au sein du laboratoire LIMATI (Laboratoire d'Innovation en Mathématiques Applications & Technologies de l'Information) de l'Université Sultan Moulay Slimane, Béni Mellal. La gestion de recherche scientifique qui représente un volet fondamental dans chaque université était parmi les premières motivations qui nous attire vers ce sujet de recherche. De plus, la recherche scientifique apparait non seulement comme un nouvel axe qui permet le développement non seulement de l'université mais également l'innovation pour l'établissement avec de nombreux avantages.

En outre, la recherche scientifique nécessite une gestion optimale qui permet d'économiser le temps de gestion et les dépenses pour les universités ce qui favorise des conditions optimales de travail. Un dirigeant doit avoir un tableau de bord qui résume le fonctionnement de la structure de recherche, centre d'étude doctorale, ainsi que tous ce qui concerne la gestion financière et administrative de la recherche scientifique à travers l'utilisation des méthodes de recherche opérationnelle. La prédiction du classement de la nouvelle publication ainsi que la prédiction du domaine de recherche du chercheur à partir de son curriculum vitæ entre dans la discipline du traitement automatique qui vise à améliorer la gestion globale de la recherche scientifique que ce soit dans son volet administrative ou pédagogique.

Pour toutes ces raisons abordées précédemment, notre vision est orientée vers l'intégration du progiciel de gestion intégré dans la recherche scientifique en vue de l'adaptabilité des besoins des dirigeants et chercheurs aux nécessités de la recherche scientifique. L'objectif principal est la proposition d'une nouvelle architecture pour la conception d'une plateforme de gestion du pôle de la recherche scientifique dans une université. Les travaux effectués ainsi que les résultats obtenus se résument comme suit :

- Fournir aux dirigeants un tableau de bord fournissant une vision globale des activités du volet de la recherche scientifique de l'université.

- Prédire le domaine de la recherche de chaque chercheur à partir du son curriculum

vitæ, en combinant la représentation numérique des textes à travers l'utilisation du traitement automatique du langage naturel et les algorithmes d'apprentissage supervisé.

- Prédire le classement des papiers scientifiques du chercheur dans leurs parcours de recherche, dans une structure de recherche, à travers un modèle prédictif basé sur l'apprentissage supervisé. Le but de notre contribution étant de recommander aux chercheurs le futur classement du papier scientifique au seins d'une base de données des papiers scientifiques similaires à leur domaine de recherche.

- Le but de la dernière contribution est de classer les structures de recherche basée sur la productivité scientifique et leur rayonnement scientifique, ainsi de fournir aux dirigeants un outil permettant d'appliquer la politique de gouvernance dans l'attribution du soutien financier pour les structures de recherche.

Problématique

La gestion de la recherche scientifique dans une université publique souffre d'un manque de système spécifique pour la recherche scientifique car la plupart des universités publiques le réduit à un simple axe dans son système de gestion globale de l'université ; ce qui prive les dirigeants d'un outil de suivi et de contrôle sur la recherche scientifique afin de promouvoir la recherche scientifique basée sur la vision stratégique pour la période 2018-2022 adopter par le CNRST ceci guide les politiques gouvernementales visant la promotion de ce secteur. D'autres systèmes utilisent des modules de gestion de recherche, mais qui reste non adapté au besoin spécifique et évolutif selon un format donné, pour pouvoir gérer les informations selon le besoin des chercheurs et des dirigeants.

Aujourd'hui, les recherches relatives aux systèmes de recherche scientifique [1] sont structurées généralement en deux groupes. Tout d'abord, on distingue celles qui sont liées aux besoins des dirigeants. Cet intérêt, qui se manifeste à tous les niveaux administratifs, est en règle générale alimenté par une volonté de rendre les savoir-faire plus opérationnels et ainsi promouvoir la promotion de la recherche par l'instauration de la transparence, cette dernière a été manifestée sur notre travail par l'utilisation de la prise de décision à critères multiples dans un but de transparence et de la bonne gouvernance pour l'attribution des subventions financiers au structures de recherche relevant de l'université. Ainsi, nous essayons à travers certaines recherches liées au traitement du langage naturel qui visent à enrichir les ressources disponibles avec des traitements automatiques des données du curriculum vitæ des chercheurs.

Ensuite, les recherches relatives aux systèmes de gestion des données propres aux

chercheurs : Dans ce type de systèmes, les connaissances sur les chercheurs sont essentielles. Nous nous intéressons aux classements des papiers scientifiques de chaque chercheur et l'apport de sa contribution scientifique dans son domaine de recherche, le chercheur doit être guidé lorsqu'il consulte ses contributions scientifiques, ce qui va lui permettre de mieux comprendre ses démarches de recherche et ainsi de pouvoir s'autoévaluer dans sa prochaine nouvelle contribution dans son domaine de recherche.

Comment pouvons-nous répondre à ces besoins en proposant aux chercheurs et aux dirigeants un parcours de recherche adapté à ces attentes en leur suggérant des méthodes qui peuvent les intéresser ?

Le travail que nous présentons dans cette thèse est une contribution pour répondre particulièrement à ces questions, puisqu'il s'intègre dans une problématique générale de management de la recherche scientifique. De plus, proposer un système qui nécessite une réflexion approfondie.

Organisation de la thèse

Cette thèse traite le management de la recherche scientifique en s'appuyant sur les progiciels de gestion intégré ainsi que l'apprentissage supervisé, le traitement du langage naturel et ces applications, le classement des papiers scientifiques et dernièrement l'utilisation des méthodes de prise de décision à critères multiples. Elle se compose de six chapitres, une introduction et une conclusion. L'introduction présente le contexte et les motivations de recherche sur le sujet de recherche scientifique ainsi que les objectifs et les contributions de cette investigation.

Le premier chapitre est réservé à la thématique de la recherche scientifique dans une université publique. Nous commençons par l'évaluation de recherche scientifique, puis nous donnons une vue générale sur les systèmes d'information scientifique, ainsi nous donnons une vision générale des contextes d'évaluation de la recherche scientifique dans quelques pays et au Maroc à travers les stratégies et les outils d'aide au pilotage de recherche et la bibliométrie.

Dans le chapitre 2, nous présentons l'historique des systèmes de progiciel de gestion intégré et les caractéristiques des environnements informatiques pour la gestion de recherche scientifique. Nous décrivons dans la suite les avantages, les enjeux de progiciel de gestion intégré, les normes et standards autour de leur utilisation. La dernière partie de ce chapitre réalise une sorte d'étude comparative entre les progiciels de gestion intégré en s'appuyant sur des critères pertinents.

Dans le chapitre 3, nous commençons par définir le système d'information, et le rôle qu'il représente pour un établissement donné, puis nous exposons notre conception du système à travers l'utilisation du UML, après nous allons aborder l'implémentation du système à travers quelque exemple d'utilisation par les chercheurs de notre université.

Les développements réalisés dans le chapitre 4 se composent de trois éléments principaux. Tout d'abord, nous introduisons le traitement automatique de la langue. Ensuite, nous exposons le processus du traitement automatique du langage naturel pour les CV. après, nous introduisons notre approche de la prédiction de domaine de recherche des chercheurs en se basant sur les différents classifieurs d'exploration de données utilisés dans le cadre de notre traitement. Enfin, nous terminons par l'évaluation de la méthode de prédiction.

Dans le chapitre 5, nous allons tout d'abord commencer par définir les processus et modèles de classement existants et les méthodes d'apprentissage et spécialement les méthodes d'apprentissage supervisés, et nous allons présenter les apports de l'intégration des méthodes d'apprentissage supervisés dans la recherche scientifique spécialement dans la prédiction du classement des papiers scientifiques.

Enfin,dans le chapitre 6, nous commençons par définir la prise de décision à critères multiples et spécialement les techniques de prise de décision multi-attributs, puis nous exposons les principes des deux méthodes de la prise de décision à critères multiples. Ensuite nous allons présenter les résultats trouvés, enfin nous comparons les résultats trouvés dans le but de découvrir la méthode la plus appropriée pour notre étude.

La thèse se termine par une conclusion générale dans laquelle nous présentons un bilan de nos travaux de recherches et nous traçons des perspectives qui nous permettraient d'améliorer ce qui a été proposé.

Chapitre 1

La recherche scientifique dans une université publique

1.1 Introduction

Le développement rapide des nouvelles technologies de l'information et de la communication a rendu le suivi de cette évolution par la recherche scientifique universitaire une nécessité très urgente. En effet, la recherche scientifique est un processus complexe qui rassemble ce qui est éducatif à ce qui est administratif et aussi ce qui est financier.
Une réflexion sur les instruments à déployer pour l'évaluation de la recherche scientifique, plus précisément pour l'aide au pilotage de la recherche scientifique dans un établissement ou une entité de recherche, à travers la méthodologie, l'accompagnement et la conception d'un système d'information, en utilisant des outils d'aide au pilotage de la recherche à la hauteur.
Dans ce chapitre, nous allons présenter en premier lieu, le contexte de l'évaluation de la recherche scientifique à l'échelle international puis le contexte marocain. Ensuite, nous allons citer les critères de l'évaluation de la recherche scientifique universitaire existant. Nous terminons ce chapitre, en introduisant la nécessité de la mise en place d'un système d'information, outil d'aide au pilotage de la recherche scientifique dans une université publique.

1.2 Le contexte d'évaluation de la recherche scientifique

Ces dernières années ont vu l'accroissement de l'intérêt porté aux rôles de la recherche scientifique dans le développement de chaque société en le considérant comme créateur de la richesse surtout dans les pays du tiers monde. En se basant sur les classements annuels des universités dans le monde. On remarque un retard très considérable pour les universités arabes et africaines (surtout les universités marocaines) ce qui influence d'une façon négative sur le rayonnement et l'effet mobilisateur que doit jouer l'université publique dans son entourage sociale et industriel. Parmi les crédibles classement donné sur

les meilleures universités dans le monde, on trouve le célèbre classement du Shanghai [2] des universités òu le Maroc est toujours absent, la même remarque pour les universités arabes à l'exception de la présence des universités saoudiennes et une université égyptienne, la même remarque pour les universités du continent africain après l'Égypte il n'y a qu'une seule université d'Afrique du sud, ce classement nous donne une idée sur l'état de la recherche scientifique dans ces établissements car parmi les critères essentiels qui sont considérés dans ce classement on trouve : la qualité de l'éducation (10%), la qualité de la faculté (40%), la publication des recherches scientifiques (40%), et la performance par personne (10%). Il s'avère que la seule exception marocain est manifestée dans l'université Cadi Ayyad qui figure au top 400 des meilleures universités mondiales classées dans le domaine des mathématiques et au top 300 dans le domaine de la physique.
Times Higher Education World University Rankings (THEWUR) [3] chargé du classement mondial des performances des universités confirme l'absence des universités marocaines dans le classement du meilleur 500 universités mondiales, la première université marocaine est classée entre 801 et 1000. Ces réalités confirment les résultats donnés par le classement du Shanghai.
La révélation de ces classements remet en question l'évaluation de la recherche scientifique que ce soit au niveau mondial, aux niveau national, et au niveau des établissements (centre d'étude doctorale, laboratoires de recherche, chercheurs), alors la mise en place des indicateurs pour évaluer la recherche scientifique est devenue une nécessité.
Les classements déjà cités donnent une idée globale sur la recherche scientifique dans chaque pays, dans la même direction, on trouve aussi les analyses bibliométriques des données communiquées des bases de données scientifiques internationales comme Scopus [4] et Science Citation Index (SCI) [5] qui sont utilisés pour l'évaluation de la recherche scientifique à travers le référencement du nombre des citations des papiers scientifiques et les journaux dans différents domaines de recherche.
Alors, la gouvernance du fonds publics investis dans la recherche scientifique dans l'université publique devient une nécessité pour les dirigeants dans l'université publique à travers la disposition des outils de pilotage stratégique, en évaluant ce qui est déjà réalisé pour mieux décider dans le future pour la valorisation de la recherche, des chercheurs et des établissements, vis à vis de leurs partenaires continentale ou internationale. Nous présenterons dans la suite une description du système d'information scientifique.

1.3 Les systèmes d'information scientifique

Les systèmes d'information scientifique [6] sont des systèmes qui visent à produire de la connaissance contrairement aux systèmes d'information d'entreprise [7] qui vise à contrôler et gérer l'activité d'entreprise.
Ainsi, ils ont été développés pour répondre aux limites du système existant en ayant recours à l'intelligence artificielle [8] pour mettre en place des systèmes plus souples et interactifs qui s'adaptent aux besoins spécifiques des chercheurs et aux dirigeants. En

outre, la recherche scientifique est devenue fortement collaborative, impliquant des scientifiques de disciplines, d'organisations et de pays différents pour faire face à la complexité des problèmes traités.
Ce système d'information rassemble des systèmes d'information distincts : la base des publications scientifiques, curriculum vitae (CV) des chercheurs, le répertoire des groupes de recherche, le répertoire des institutions de l'université, ainsi que la base du soutien de recherche scientifique, etc.

1.4 Les contextes d'évaluation de la recherche scientifique

De manière générale, des procédés d'évaluation de la recherche ont été instaurés par plusieurs pays qui ont une crédibilité et avancement dans la scène internationale :
- En Royaume Uni : met en œuvre des évaluations périodiques pour l'attribution des subventions par Higher Education Funding Councils (HEFC).
- En Italie : met en œuvre des évaluations pour l'attribution des financements par la création de Italian National Agency for the Evaluation of the University and Research Systems(ANVUR).
- En Espagne : certifie et accrédite et évalue par National Agency for Quality Assessment and Accreditation (ANECA).
- En Belgique : une Agence pour l'Évaluation de la Qualité de l'Enseignement Supérieur (AEQES) a été créée dans la communauté française et un système d'assurance qualité a été introduit pour la communauté flamande.
- En Allemagne : l'initiative d'excellence incite la recherche de pointe et l'amélioration significative de la qualité des installations de recherche et d'enseignement supérieur allemandes visant un surcroît de financement aux établissements.
- En Etats-Unis : une évaluation est réalisée pendant le processus d'accréditation des établissements.
- En Chine : le développement d'évaluations prenant la forme de classements consultables sur internet par exemple le classement des universités du Shanghai.
- En France :la création d'Agence d'Évaluation de la Recherche et de l'Enseignement Supérieur (AERES) dans le but de pratiquer une évaluation intégrée, accompagnées des observations des entités évaluées.

1.5 Le contexte Marocain

Au Maroc, un retard considérable a été remarqué concernant la création d'un organisme étatique spécialisé dans l'évaluation des recherches scientifiques, puis la création au milieu des années soixante-dix du Centre National de Coordination et de Planification de la Recherche Scientifique et Technique (CNCPRST) qui a pour mission d'orienter et de coordonner les recherches scientifiques et techniques de tous ordres, après la mise en place du centre national pour la recherche scientifique et technique (CNRST) qui a pour

objectif de :
- Mettre en œvre des programmes de recherche et de développement technologique.
- Contribuer à la diffusion de la publication de travaux de recherche.
- Apporter son concours au renforcement de l'infrastructure nationale de recherche.
- Contribuer à la valorisation et au transfert des résultats de recherche.
- Etablir des conventions dans le cadre des activités de recherche.
- Créer des synergies entre les différentes équipes de recherche qui travaillent sur des thématiques prioritaires (réseaux, pôles de compétence).
- Evaluer et faire le suivi de toutes les activités de recherche.
Bien que de nombreux actions ont été menées par la CNRST, d'autres initiatives ont été prises comme :
- La création du fond national du financement de la recherche scientifique et du développement technologique.
- La création du comité interministériel permanent de la recherche scientifique, de l'innovation et du développement technologique.
- L'organisation et la structuration du système national de la recherche et d'innovation avec ses différentes composantes matérielles, humaines et légales à l'échelle nationale.
- La fixation des priorités nationales en matière de recherche scientifique.
Dans cette perspective, la CNRST a élaboré une vision stratégique pour la période 2018-2022 afin de renforcer son positionnement dans le système national de la recherche et de l'innovation qui comprend six axes stratégiques :
- Renforcer les mécanismes de bonne gouvernance.
- Soutenir et financer la recherche scientifique et encourager l'excellence.
- Renforcer le système national d'évaluation des résultats de la recherche et de l'innovation.
- Promouvoir les synergies et encourager la mutualisation.
- Renforcer le partenariat international et la coopération dans le domaine de la recherche scientifique.
- Contribuer au rayonnement de la recherche scientifique nationale et améliorer sa visibilité.

Malgré les efforts réalisés par le CNRST, un manque dans l'évaluation et l'accréditation reste remarqué, ce qui a conduit à l'apparition d'Agence Nationale d'Évaluation et d'Assurance Qualité de l'Enseignement Supérieur et de la Recherche Scientifique (ANEAQ) qui vient combler ce vide .Cette agence a réalisé des partenariats avec d'autres agences avec les mêmes orientations comme le Réseau Arabe pour l'Assurance Qualité dans l'Enseignement Supérieur (ANQAHE) et le Réseau Francophone des Agences Qualité pour l'Enseignement Supérieur(FrAQ-Sup), parmi les missions principale de l'agence, nous citons :
- L'évaluation des activités des centres d'études doctorales.
- Dresser le bilan des formations et des travaux de recherche réalisés dans ces centres.

- L'évaluation de la recherche scientifique et l'efficacité de ses structures.
- L'évaluation des programmes et des projets de coopération universitaire dans le domaine de la formation et de la recherche scientifique.

Un rapport d'activité et d'évaluation et de suivi 2016-2017 d'un échantillon de filières à accès régulier accréditées au titre de la session 2017 a été publié jusqu'à maintenant. Dans ce rapport nous ne trouvons pas d'évaluation concernant la recherche scientifique dans les universités publiques, seulement des audites des filières et des établissements.
Dans cette situation, les classements internationaux comme le classement de Shanghai et THE, nous donne une vision réaliste sur l'état de la recherche scientifique dans nos universités publiques car le principe d'évaluation en place reste qualitative contrairement à ces nouveaux modèles de classement qui sont plus attirant pour la simple audience.
D'autres moyens peuvent nous donner une idée sur la recherche scientifique comme les bases de données scientifiques reconnues comme (ISI Thomson, Scopus,...) qui fournit des statistiques sur le nombre de papiers publiés par chaque université par années, et aussi les journaux qui relèvent de certaine entité de recherche.
Donc, en se basant sur les classements internationaux déjà cités et les statistiques fournis par les bases de données internationaux, un retard remarquable par rapport au niveau international est noté, alors la mise en place d'une évaluation est nécessaire à partir de la structure de recherche en passant par les centres des études doctoraux puis les établissements et finalement par la ministère de l'éducation nationale, de la formation professionnelle, de l'enseignement supérieur et de la recherche scientifique.
Alors, la mise en place d'un système d'information universitaire focalisant sur la recherche scientifique au niveau de chaque université publique permet aux dirigeants au niveau de chaque université de bien détecter les anomalies concernant la recherche scientifique au niveau de chaque établissement ; et au niveau national la nécessité de la création d'une base de donnée nationale commune concentrant tous les données recueillis de chaque université comme l'exemple de la base de donnée Pascal & Francis en France [9], pour obtenir un tableau de bord au niveau national qui sera une base pour tous les systèmes d'évaluation ou les rapports générés par les commissions ou les agences que ce soit publique ou privée qui auront la mission de détecter le disfonctionnement dans notre système de recherche scientifique universitaire.
Actuellement, après l'adaptation de la carte universitaire au nouveau découpage régional, ce dernier permet de regrouper un ensemble d'établissement. Ceci va permettre d'avoir un potentiel de recherche important. Alors, la mise en place d'un tableau de bord national de recherche scientifique pour les dirigeants, représentera un instrument d'aide à la gouvernance et au pilotage et assurera sa visibilité et son rayonnement par rapport à son environnement de recherche internationale.

1.6 La gestion stratégique de la recherche publique

Parmi les stratégies nationale adoptées au niveau de la recherche scientifique on trouve la stratégie nationale pour le développement de la recherche scientifique à l'horizon 2025 qui proposera quelques mesures parmi lesquelles :
- Rendre attractif le métier de chercheur.
- Instaurer un statut de chercheur pour les personnes qui exercent une activité de recherche dans des établissements de recherche, sans être des enseignants-chercheurs.
- Dégager les ressources nécessaires pour pérenniser les structures de recherche.
- Systématiser les évaluations et en indexer la carrière sur la production scientifique, dans le but de faire évoluer les structures actuelles répondant aux normes internationales.
- Intensifier, diversifier, faciliter les échanges scientifiques et renforcer les réseaux scientifiques existants.
- Veiller à la mise en place de bibliothèques et d'une centrale des thèses, sources documentaires écrites et électroniques, accessibles à l'ensemble des chercheurs, avec un accès aux bases de données des départements ministériels (statistiques, intérieur, etc.).
- Organiser et promouvoir la publication scientifique en regroupant les publications par grandes familles de disciplines à l'échelle universitaire afin de remédier à la dispersion actuelle des efforts.
- Prévoir un accroissement de la part du PIB consacrée à la recherche et l'innovation qui devra atteindre 3% à la place de 1% maintenant.
- Mettre en place des mécanismes permettant la mobilité des acteurs de la recherche (enseignants chercheurs, chercheurs, ingénieurs, médecins, cadres, etc.) entre les universités, les instituts et le monde socio-économique.
- Alléger et assouplir les procédures de gestion financière des budgets de recherche soit au niveau du ministère ou au niveau des établissements.
- Augmenter les moyens financiers alloués à la recherche et assurer une répartition plus équilibrée entre champs disciplinaires.
- Poursuivre et approfondir la politique de coopération initiée depuis ces dernières années,
- Définir des choix et des orientations des projets de recherche.
- Développer de la culture d'entreprenariat dans les milieux académiques pour permettre à certains éléments de cette population de jouer un rôle actif dans la création d'entreprises innovantes basées sur la valorisation des résultats de la recherche.

La réalisation de ces ensembles de mesure nécessite une commission d'experts, qui veille sur la mise en place des outils d'aide à la décision qui facilite l'évaluation permanente de la recherche scientifique au seins des universités publique.

Dans

1.7 La bibliométrie au service de l'évaluation scientifique

La scientométrie [10] considérée comme la science de la mesure et l'analyse de la science qui se base sur un ensemble d'entrée et un ensemble de sortie recours dans le domaine d'étude des publications à la bibliométrie [11], qui est une méta-science qui prend la science pour objet d'étude en se basant sur trois éléments de l'activité scientifique : ses intrants, ses extrants et ses impacts.
La publication scientifique représente toutes les publications dans des journaux ou des conférences soit des chapitres de book scientifiques ou des brevets scientifiques. Tous ces types de publications représentent le fruit de travail d'un chercheur qui publie ces travaux dans le but de faire circuler ces résultats dans des bases de données qui ont une large visibilité internationale et crédibilité scientifique comme : (ISI Thomson Reuters, Scopus,) et des maisons d'édition de renommé comme (Elsevier, Springer, Wiley Online Library,) ,mais avec tous ces effort fournis, les avantages qu'on peut tirer de ces publications restent limitées si on ne peut pas gérer cette masse de publication qui vient s'ajouter chaque jour au millier ou des millions de papiers scientifiques existants ; d'oû vient la nécessité de trouver un compromis entre les différentes bases de données pour essayer de créer des méta données sur chaque publication scientifique [12] qui contient :le titre les auteurs, l'affiliation de chaque auteur, résumé, les mots clés, type de papier, journal de publication, date de publication, discipline de papier.
Tous ces données et d'autres représentent l'essentiel des données fournis pour chaque papier par les bases de données qui permettent pour la bibliométrie de réaliser des traitements statistiques, à travers l'application de quelques lois et théories en relation comme la théorie de l'information, la loi de Pareto, la loi Binomiale et d'autres.
Pour procéder à un traitement automatique, la bibliométrie nécessite la présence de données structurés appelés métadonnées organisées sur les papiers scientifiques ce qui doit être normalisé [13], qui dit comparé des études bibliométriques dit avoir une base de données commune à partir duquel on réalise des études, ce qui n'est pas le cas ici au Maroc oû on trouve que la CNRST à travers son projet Eressources IMIST a mis en place une plateforme qui permet l'accès à l'ensemble des bases de données.

1.8 Les indicateurs au service de l'évaluation scientifique

Celui qui s'intéresse à l'évaluation a besoin des indicateurs pertinents dans n'importe quel domaine. Dans le cas dans la recherche scientifique, pour que les dirigeants évaluent les besoins, ils devront avoir un tableau de bord qui permet d'avoir une idée globale sur le fonctionnement de ces structures de recherche à l'échelle du structure de recherche pour les dirigeants au niveau d'établissement, ils doivent avoir une idée sur la recherche scientifique au niveau de chaque établissement relevant de l'université pour les dirigeants au niveau de l'université, ainsi avoir une idée au niveau de la recherche des universités pour les dirigeants au niveau du ministère de l'éducation nationale, de la formation

professionnelle, de l'enseignement supérieur et de la recherche scientifique.
Dans le cas d'une structure de recherche relevant d'établissement :
- Nombre de nouveaux étudiants inscrits chaque année.
- Nombre de papiers scientifiques publiées chaque année.
- Nombre de thèses de doctorat soutenues chaque année.

Dans le cas de chaque établissement relevant de l'université :
- Nombre de nouveaux étudiants inscrits par structure de recherche.
- Nombre de papiers scientifiques publiées par chaque structure de recherche par année.
- Nombre de thèses de doctorat soutenues par chaque structure par année.
- La structure la plus productive.
- Le nombre des mobilités des chercheurs accordés par structure de recherche par année.
- Le nombre d'événements scientifiques organisés par structure de recherche par année.

Dans le cas de chaque université publique :
- Nombre de nouveaux étudiants inscrits par établissement relevant de l'université.
- Nombre de papiers scientifiques publiées par établissement par année.
- Nombre de thèses de doctorat soutenues par établissement par année.
- La structure la plus productive par établissement.
- Le nombre des mobilités des chercheurs accordés par établissement.
- Le nombre d'événement scientifique organisés par établissement.

Concernant les indicateurs utilisés par Scopus on trouve :
- H-index : est basé sur le nombre le plus élevé d'articles ayant au moins le même nombre de citations.
- CiteScore : mesure la moyenne des citations reçues par document publié dans la publication en série.
- SJR : mesure les citations pondérées reçues par le périodique, la pondération des citations dépend du domaine et du prestige de la série citant.
- SNIP : l'impact normalisé par papier de la source qui mesure les citations réelles reçues par rapport aux citations attendues pour le domaine de la publication en série.

Concernant les indicateurs utilisés par ISI Web of Science nous trouvons :
- H-Index : l'indicateur de recherche le plus utilisé qui mesure à la fois la productivité et l'impact de la production scientifique d'un auteur.
- Le facteur d'impact : mesure l'importance d'une revue en fonction du nombre de citations reçues dans une année.
- Journal Citation Reports : produit de ISI Web of Knowledge et une ressource faisant autorité pour les données de facteurs d'impact.

Nous trouvons plusieurs d'autres indicateurs [14] qui servent à donner une vision sur

les auteurs et les journaux de chaque base de données, l'étude des différents indicateurs des journaux et de la bibliométrie constitue un axe de recherche en plein développement et dynamisme. Il existe aussi d'autres indicateurs descriptifs, qui ne concernent qu'une entité (nombre d'articles d'une revue, d'un auteur, d'une structure, nombre de références,...), les indicateurs relationnels (nombre de citations, ...).
Parmi tout cet ensemble des indicateurs que ce soit pour les structures de recherche ou pour les base de données ou les journaux ou les auteurs, chaque communauté scientifique a ces propres besoins et ses propres pratiques et perceptions de la recherche scientifique ce qui nous fournissent les indicateurs spécialisés ou locales, en revanche, l'application de certains indicateurs nécessite des bases de données avec d'énorme données pour fournir des indicateurs fiables, ce qui entraine un disfonctionnement de certain indicateur qui nécessite des masses de données considérables.
Mais la plupart de ces indicateurs rencontre plusieurs critiques car tout d'abord :
- Il ne vise pas la qualité en se focalisant sur la quantité de productions scientifiques, ce qui entraine plusieurs forme d'anomalie qui dégrade la qualité de la production scientifique comme en essayant de publier un grand nombre de papiers scientifiques, on découpe une idée d'un article sur plusieurs : publié des travaux de recherches qui reprend le même travail dans plusieurs forme en essayant d'être cité par d'autres chercheurs ou en incitant les chercheurs a référencer des papiers scientifiques afin d'avoir de bon résultat pour certain indicateur comme le H-Index ou d'autre.
- Concernant les bases de données scientifiques comme Scopus ou WOS, ces dernières favorisent les travaux anglophones et les journaux anglophones ce qui construit un obstacle réel pour les chercheurs de certains pays comme le nôtre, ce qui diminue la visibilité de la recherche scientifique comme nous avons déjà cité dans le classement du Shanghai, et d'autres.
- La publication des travaux de la recherche dans une conférence ou workshop n'est pas prise en compte par certain indicateur.
- Les bases de données scientifiques favorisent l'apparition des travaux scientifiques les plus cités en ignorant les travaux récents en s'appuyant sur des indicateurs qui se basent sur le nombre de citation pour chaque production scientifique.

Le Centre national pour la recherche scientifique et technique (CNRST) met en place un programme de soutien à la recherche multidisciplinaire dans les domaines en relation avec la pandémie actuelle du « Covid-19 » pour faire face à la pandémie du « Covid-19 », et afin de mieux se positionner sur la scéne internationale en terme de production scientifique dans les bases de données international les plus renommées ,vient notre publication du travail intitulé « Bibliometric method for mapping the state of the art of scientific production in Covid-19 ». (voir annexe)

1.9 Les outils d'aide au pilotage de la recherche scientifique

L'universités marocaine publique s'approprie de plusieurs logiciels pour sa gestion intérieure attribuée par le ministère de l'éducation nationale, de la formation professionnelle, de l'enseignement supérieur et de la recherche scientifique soit pour la gestion des étudiants, des ressources humaines, de ces appels d'offres, ... Mais pour la gestion de la recherche scientifique, chaque université essaye avec ces propre moyen de gérer ce pilier primordial soit du côté financement des structures de recherches, ou pour les programmes de mobilité, l'organisation des évènements scientifiques, gestion de la production scientifique... D'où vient le développement et l'adoption de l'application SIMarech [136] par quelques universités mais cette dernière ne couvre pas la totalité des axes relatives au pilier de la recherche scientifique ; donc l'adoption d'un système qui permet de gérer ce pilier constitue une nécessité pour tous les intervenants à partir des chercheurs en passant par les structures de recherches et les centres d'études doctorales au chefs d'établissements, en arrivant au sommet de l'université le président pour avoir une vision sur le bon fonctionnement de la recherche scientifique.
Parmi les avantages attendus du système :
- Faciliter la tâche du dirigent en fournissant un tableau de bord qui permet une maîtrise parfaite de processus du développement de la recherche scientifique.
- Gérer le processus de préinscription en thèse, d'inscription en thèse, réinscription, et dépôt thèse pour la soutenance.
- Gouverner le pilier de la recherche scientifique dans l'université.
- Gérer les coopérations de l'université.
- Valoriser la recherche à travers l'exploitation des données collectées.
- Faciliter l'échange entre les chercheurs à travers la plateforme de discussion.

La réalisation de ce système nécessite la réalisation d'une base de donnée commun à tous les établissements de l'université qui permette au tableau de bord de recharger ces informations en permanences pour assurer les taches suivantes :
- Traitement des données collectés.
- Accessibilité par tous les établissements.
- Mise à jour des indicateurs sur le tableau de bord.
- Accessibilité par les chercheurs, les chefs des structures de recherches, les doyens, les présidents des universités.

1.10 Conclusion

Dans ce chapitre, nous avons présenté une vision globale sur la recherche scientifique dans une université dont le but est d'utiliser des technologies de l'information et de la communication pour faciliter et améliorer la qualité des services fournis aux chercheurs. Ainsi, nous avons expliqué le contexte de l'évaluation de la recherche scientifique. En

outre, nous avons mis en évidence le rôle des systèmes d'information scientifique qui vise principalement à produire de la connaissance.

Cependant, nous avons essayé de fournir une étude détaillée des contextes d'évaluation de la recherche scientifique internationale dont le but est de se situer par rapport à notre propre contexte marocain de recherche scientifique. Un bref état de l'art de l'existant est exposé pour avoir une idée claire sur l'actualité de la recherche scientifique et les différents rapports et programmes adoptés par la ministère tutelle à travers ces agences et centres dans ce domaine, ainsi que les stratégies adoptées pour mettre la recherche scientifique publique marocaine dans la bonne direction afin que la recherche scientifique soit compétitive au niveau régional et international.

Le rôle de la bibliométrie en relation avec l'évaluation scientifique a été abordé, à travers les différentes bases de données scientifique, et les indicateurs qu'on peut tirer concernant les chercheurs, les journaux, et les indicateurs locaux ou nationaux, statistiques ou dynamiques.

En résumé, la recherche scientifique a connu une évolution croissante au cours de ces dernières années, les travaux de recherches en cours tentent de résoudre les problématiques d'évaluation en s'appuyant sur plusieurs méthodes et mesures afin d'aider les dirigeants au niveau de l'université ou le ministère tutelle.

Chapitre 2

Progiciel de gestion intégré

2.1 Introduction

L'évolution des établissements [16] en raison du changement rapide des technologies de l'information et de la communication, ainsi, le changement continue des besoins constitue un facteur de changement et dynamisme continu. L'objectif est de réaliser une sorte d'équilibre entre ces différents processus et mécanismes à travers la réalisation des formations et accompagnement du changement pour les différents intervenants.
Les établissements sont de plus en plus dépendants des technologies de l'information et de communication qui se développent de jour en jour et deviennent plus dispensable, ce qui rend leur utilisation par les établissements une nécessité mais en revanche, il demande du budget et de l'expertise pour le personnel. Ce qui constitue un risque pour l'adoption et l'implémentation de ces dernières au sein de l'établissement [17] .
Cependant, les progiciel de gestion intégré [18] sont actuellement fréquemment utilisés par les établissements, car il fournissent la majorité des composantes fonctionnelles (comptabilité, ressource humaines ...) dans l'établissement a besoin, ainsi à travers l'utilisation d'une base de données unique pour la gestion de toutes ces composantes. En dépit de ces avantages déjà cités, l'adoption d'un progiciel de gestion intégré nécessite l'implication de tous les services de l'établissement, le coût d'implémentation ainsi que la résistance au changement par le personnel à tous les niveaux. Ceci, nécessite un effort au niveau de formation et d'accompagnement par des experts à la matière [19].
En outre, il faut souligner que les progiciels de gestion intégrés sont des méthodes de planification des besoins en composantes prometteuses pour les établissements. Vu qu'ils sont en pleine expansion durant ces dernières années. Le développement des systèmes d'informations des établissements peut bénéficier de cette vue globale du progiciel de gestion intégré, ou toutes les ressources sont disponibles et adaptés à la gestion globale de l'établissement.
Dans ce qui suit, nous présentons d'abord un bref historique des générations du progiciel de gestion intégré [20]. Ensuite nous enchainons avec l'architecture du PGI, après, nous discutons les avantages d'utilisation des PGI pour l'université.

2.2 Historique du progiciel de gestion intégré

Cette dernière décennie a connu une évolution considérable des technologies de l'information et de la communication, une évolution marquée par la croissance permanente des besoins des établissements en terme de numérisation et de la digitalisation à travers ces novelle technologies. Pour mieux comprendre les enjeux et les différentes phases de l'évolution du progiciel de gestion intégré , nous présentons dans la suite une synthèse des différentes générations du PGI [21], qui devrait nous fournir quelques éléments clés de compréhension du rôle du PGI pour un établissement donné (Figure 2.1).

Figure 2.1 – *Evolution du Progiciel de Gestion Intégré.*

- 1960 : Systèmes juste pour la gestion des stocks et contrôles.
- 1970 : Planification des besoins en matériel (MRP) [22], marqué par le début d'utilisation de l'informatique pour automatiser les procédures ; chaque service avait son propre système d'information et ces applications étaient développées indépendamment les unes des autres ce qui entrainera des problèmes aux établissements. C'est pour cette raison qu'a vu le jour MRP pour résoudre ces problèmes.
- 1980 : Planification des ressources de production (MRP II)[23], soutiendra les efforts visant à optimiser les processus de fabrication en synchronisant les matériaux avec les exigences de production.
- 1990 : Continuité du web pour les progiciels de gestion intégré(PGI) qui devient un standard dans les entreprises. Il adopte une solution standardisée pour tous les services et centralise tous les données dans une base de donnée unique.
- 2000 : Progiciel de gestion intégré (PGI) s'étend à la gestion de la relation client et offre des services étendus tels que le commerce électronique, l'entreposage, la logistique, la planification des capacités.
- PGI II : Conservation simplifiée des données pendant la synchronisation et réduction du nombre de logiciels requis au sein des grandes entreprises.
- PGI III : L'intégration des clients dans les systèmes (PGI) constitue l'essence du concept (PGI III). Il peut être adapté par la mise en œuvre des dernières réalisations informatiques

les plus sophistiquées, en se basant sur la mise en œuvre de l'informatique en nuage et les technologies GRID dans les systèmes PGI.

2.3 Principes de base d'un PGI

2.3.1 Serveur PGI

Un serveur PGI est un système similaire à un ordinateur. On y stocke l'ensemble des informations pour les utilisateurs des différents modules PGI. Ce serveur est au cœur de l'architecture PGI puisqu'il est utilisé pour :
- Administrer le réseau.
- Gérer les connexions des différents utilisateurs.
- Mutualiser les informations.
- Assurer la traçabilité.

2.3.2 Modules PGI

Un module PGI est une fonctionnalité du PGI spécialisée dans la gestion d'une activité de l'établissement. Ce module rassemble un ensemble de fonctionnalités :
- Compatibles avec les autres modules.
- Connectées en permanence à la base de données commune.
- Mutualisées et actualisées en temps réel.

Les modules PGI les plus courants sont les suivants :
- Gestion : des achats, des ventes, ...
- Comptabilité : classique, clients, fournisseurs.
- Contrôle de gestion.
- Production.
- Organisation du travail.
- Stockage, archivage, inventaire.
- Ressources humaines (payement, congés, pointage...).

2.3.3 Architecture du PGI

L'architecture technique est principalement client / serveur au trois niveaux [24], comme décrit dans la (Fig.2.2).

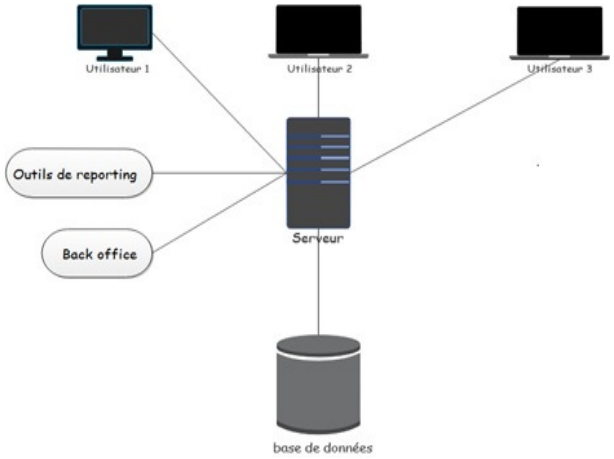

FIGURE 2.2 – *Evolution du Progiciel de Gestion Intégré.*

- Le niveau « Présentation » : il constitue l'interface utilisateur.
- Le niveau « Applications » : il correspond aux fonctions de traitement de l'information.
- Le niveau « Base de données » : il gère les grands volumes de données que l'établissement conserve avec une base de données unique et disponible pour tous les utilisateurs.

L'architecture d'un PGI constitué principalement par un serveur avec une base de données unique disponible pour tous les utilisateurs. Ce qui permet l'utilisation de différents réseaux intranet, extranet. De plus, un PGI est constitué d'un ensemble de modules qui peuvent fonctionner les uns avec les autres avec des possibilités de :
- L'utilisation d'une base d'information unique.
- La mise en réseau de différentes stations.
- La compatibilité garantie entre les différents modules.

2.4 Caractéristiques d'un PGI

Parmi les caractéristiques d'un PGI :
- Système flexible qui répond aux besoins d'un établissement.
- Couvrir l'ensemble du système d'information (SI) de l'établissement.
- Garantir le caractère unique des informations.
- Prendre en charge plusieurs plates-formes matérielles pour les établissements ayant une collection hétérogène des systèmes.
- Accélérer l'optimisation des processus de gestion.
- Partager la même base de données.
- Favoriser la communication interne et garantir l'intégrité et le caractère unique du système informatique.

Un des plus importants avantages d'un PGI, est la mise à jour de données en temps réel et sa transformation aux modules cibles, sans interfaces entre les modules, ainsi il

garantit une synchronisation du processus.
En d'autres termes, un système PGI relie les informations en utilisant des données communes capables de transformer les données transactionnelles en informations utiles. De plus, un PGI doit avoir une collection des meilleurs processus d'affaires applicable dans le monde entier pour répondre aux différents besoins hétérogènes de l'établissement.
De plus, un progiciel PGI impose sa propre logique, stratégie, culture et organisation à l'établissement. Certes, les fonctionnalités majeures du PGI sont fournis en multi-plateforme, multi-installations, fabrication multimode, multidevises et multilingues qui couvrent tous les domaines fonctionnels comme la fabrication, la vente et la distribution, les dettes, les créances, l'inventaire, les comptes, les ressources humaines et les achats, ...

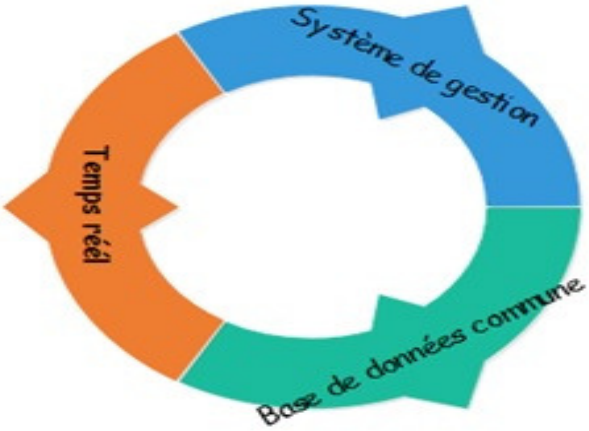

FIGURE 2.3 – *Caractéristiques d'un système PGI.*

2.5 PÉRIMÈTRE DE GESTION D'UN PGI ?

De nombreuses établissements choisissent de mettre en œuvre un système PGI afin de bénéficier de plusieurs avantages, tels que la réduction des coû ts d'exploitation, l'augmentation de la productivité et l'amélioration des services.
Le but d'un système PGI est d'homogénéiser le système d'information de l'établissement avec un outil unique, capable de couvrir un large périmètre de gestion, tel que :
- Gestion des achats.
- Gestion des affaires.
- Comptabilité.
- Contrôle de gestion.
- Gestion des stocks, logistique, transport ....

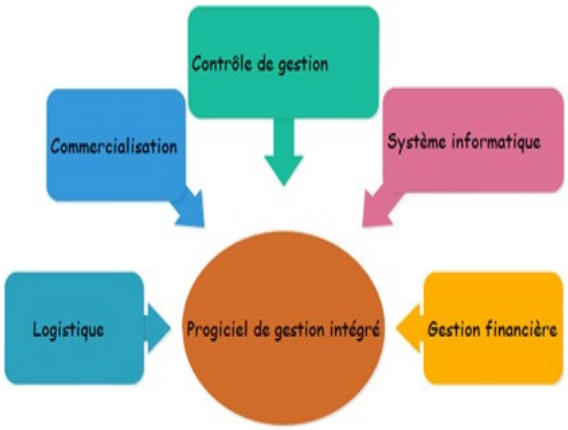

FIGURE 2.4 – *Secteur d'activité du PGI.*

2.6 LES AVANTAGES ET LES INCONVÉNIENTS DES PGI

2.6.1 Les avantages du PGI

Les PGI ont beaucoup d'avantages pour les établissements [25], parmi ces avantages nous citons :
- L'unicité du système d'information pour éviter la redondance d'informations entre différents (SI) d'établissement.
- L'intégrité du système d'information.
- La communication interne et externe.
- La cohérence et homogénéité des informations.
- La globalisation de la formation pour les employés qui doivent apprendre le fonctionnement d'un seul logiciel.
- La meilleure coordination entre les différents services.
- La création d'un environnement de travail standardisé, identique pour tous.
- L'optimisation des processus.
- La diminution des coûts.
- L'utilisation évolutive.
- La productivité.
- L'unicité : l'établissement n'est plus obligé d'utiliser différents programmes pour gérer les multiples services.
- La personnalisation : l'autre atout du PGI est sa capacité d'adaptation aux besoins.

2.6.2 Les inconvénients du PGI

Parmi les inconvénients du PGI [26] nous citons :
- La complexité : avant de déployer un PGI, l'établissement doit avoir connaissance de l'ensemble de ses processus et de leur fonctionnement. Sinon, le périmètre couvert par le PGI ne sera pas total et son efficacité pourra être réduite.
- Le coût : en général, le coût lié au déploiement du PGI et sa maintenance sont élevés.

- La dépendance envers l'éditeur du PGI : il est rare, en pratique, de changer de PGI une fois qu'il a été déployé. L'établissement doit s'assurer qu'elle fait le bon choix, compte tenu de ses besoins.
- Le matériel adéquat : la base de données étant volumineuse, un PGI nécessite l'installation du serveur puissant.
- La nécessité d'une maintenance continue.

2.7 Les différents types de PGI

Aujourd'hui, la gestion des établissements nécessite l'utilisation de plusieurs logiciels qui sont très hétérogènes. Afin de faciliter et d'optimiser les différentes tâches des établissements, ces derniers ont opté de travailler avec le système PGI. C'est pourquoi il est essentiel de procéder à une étude entre PGI payant et le PGI open source [27].

2.7.1 PGI propriétaire

Un PGI propriétaire est un progiciel créé par une société spécialisée dans la conception et la mise en place de logiciels et de systèmes informatiques. Choisir un PGI propriétaire c'est profiter :
- D'un savoir-faire reconnu.
- D'un accompagnement à toutes les étapes du projet.
- D'un service dédié assurant l'étude, la mise en place, la maintenance et le service après-vente.
- D'un service personnalisé adapté à l'activité de l'établissement.

Parmi les principaux PGI propriétaires du marché, nous pouvons citer :
- SAP (leader mondial).
- ORACLE/PEOPLESOFT.
- SAGE ADONIX.
- MICROSOFT.
- SSA GLOBAL.
- GEAC.
- INTENTIA/LAWSON.
- INFOR GLOBAL SOLUTIONS.

2.7.2 PGI open source

Un PGI open source est un logiciel libre et donc moins cher puisqu'il n'implique pas l'acquisition d'une licence. Ce qui rend ce dernier très prisé par les petites ou moyennes entreprises [28] , qui souhaitent profiter de tous les avantages d'un PGI à moindre frais. Choisir un PGI open source possède de nombreux avantages :
- Logiciel libre : il ne nécessite pas l'acquisition d'une licence, ce qui permet de faire de sérieuses économies, et moins cher qu'un PGI propriétaire.

- Absence de licence sur PGI open source. Ceci donne une forme d'indépendance aux établissements qui ne prennent aucun engagement.

Parmi les PGI open source du marché nous trouvons :
- ARIA.
- OPEN ERP/ODOO.
- COMPIERE.
- FISTERRA.
- OFBIZ.
- OPENBRAVO.
- PGI SUITE.
- TIOLIVE.

2.8 Etude comparative

Les systèmes PGI constituent la plus grande application logicielle adoptée par les établissements, et leur mise en œuvre nécessite des montants importants [29]. Cependant, quelques recherches ont été menées sur l'utilisation du PGI dans un environnement universitaire [30], par rapport à d'autres environnements. Les universités diffèrent d'autres organisations et entreprises en possédant certaines spécificités [31], des circonstances différentes ; environnement de travail différent utilisant les PGI à des fins académiques et non lucratif comme les entreprises. Parmi les raisons qui poussent les universités à choisir un système PGI open source :
- Moins cher : les systèmes PGI open source minimisent les coûts d'organisation des entreprises à long terme, aucune licence n'est nécessaire. De plus, sa mise en œuvre est librement disponible et économique par rapport aux systèmes PGI commerciaux.
- Flexibilité : les systèmes PGI open source sont plus flexibles que d'autres systèmes. Lorsque nous mettons en œuvre un PGI Open Source, de nombreuses nouvelles interfaces seront créées. De plus, cette flexibilité donne au PGI open source une mise à niveau facile vers une nouvelle version sans difficultés et sans perte d'informations.
- Contrôle total : les organisations ont le contrôle total sur le système et peuvent modifier la source. De plus, chaque organisation peut utiliser un PGI open source en fonction de ses propres besoins.
- Efficacité : l'utilisation du PGI par les organisations améliore l'efficacité, car elle élimine une partie du processus manuel existant. Le stockage et l'accès aux données deviennent plus efficaces et de nombreux processus sont intégrés.
- Meilleur rapport : les PGI améliores les rapports. Cette option est importante pour l'organisation car plusieurs rapports doivent être générés.
- Sécurité des données : l'une des exigences du PGI est d'utiliser des données précises et pertinentes.

D'après des études réalisées par Smile [32] une société du benchmarking entre les PGI open source dominant le marché selon les six critères suivants :
- Notoriété actuelle : sont considérés :
  — Nombre et importances des références clients.
  — Nombre et notoriété des intégrateurs existants (s'agit-il uniquement d'amateurs isolés ou de vraies entreprises ? est-ce qu'il n'y a qu'un seul intégrateur derrière un projet ?
- Dynamique : il s'agit d'une dynamique communautaire autour de la solution open source, avec la qualité technique, elle va déterminer directement la place de la solution dans le futur. Cette dernière prend en considération :
  — La gouvernance : dans quelle mesure les intégrateurs et les utilisateurs sont-ils consultés concernant la conception et l'évolution du produit ?
  — La fréquence des mises à jour de la documentation, notamment des wikis.
- Technologie : investissements et communauté sont encore peu de chose devant la cohérence, la puissance et l'adéquation avec les standards des modélisations au cœur d'un PGI. Cette dernière prend en considération :
  — Le respect des standards existants.
  — La puissance et canonicité des abstractions mises en jeu.
  — Le degré de factorisation du code.
  — La maturité et couverture des web services.
    - La modularité de l'application.
    - L'absence de problème de performance.
- Périmètre : Il s'agit ici du volume global des fonctionnalités. Beaucoup de ces dernières ne sont jamais utilisées ou devront être modifiées. D'autant plus que sur un PGI souple, l'ajout d'une fonctionnalité peut se révéler relativement simple.
- Souplesse : Dans la mesure où on doit souvent dépasser le périmètre fonctionnel natif de l'outil, Il s'agit donc d'un critère absolument déterminent dans le coût total de possession compte-tenu du fort coût relatif des développements spécifiques. La souplesse rejoint ici la technologie mais elle met spécifiquement l'accent sur la modularité de la plateforme du PGI et sur l'efficacité du développement par des tierces parties.
- Ressources : Les PGI étudiés ont une très bonne capacité à être configurés et requièrent donc moins de développement spécifique.
Le tableau qui suit résume cinq PGI open source selon les six critères vus précédemment en leur fixant à chacun une note allant de 0 (faible) à 5 (excellent).

	Notoriété	Dynamique	Technologie	Périmètre	Souplesse	Ressources
OPEN ERP/ODOO	4	5	4	5	5	4
OPENBRAVO	4	5	3	4	3	4
COMPIERE	5	3	3	4	3	4

TABLE 2.1 – *Tableau comparatif des principaux PGI Open source.*

CONCLUSION

Les PGI sont très connus et ont pour objectif de fournir des solutions génériques qui servent les différents établissements. Selon les chiffres sur le marché mondial, ce produit est un succès, il offre une grande stabilité pour les établissements.
Dans ce chapitre, nous avons vu que le principal rôle d'un PGI, est de répondre aux attentes opérationnelles et informationnelles des responsables des établissements, des attentes formulées sous forme de besoins venant de leurs directions internes ou des utilisateurs externes. D'autre part les PGI ont pallié aux problèmes des installations hétérogènes, qui a causé un énorme souci par le travail avec des systèmes d'informations composés de nombreuse applications qui ne communiquent pas forcément entre elles.
Le choix d'un PGI dans un établissement, n'est pas une tâche aussi simple et confortable pour les dirigeants, car il relève de l'avenir organisationnel et gestionnaire du secteur d'activité de l'établissement.
Dans ce chapitre, nous avons mis l'accent sur le potentiel du PGI comme une solution adéquate pour implémenter un système de management de recherche scientifique, du moment qu'il fournit tous les moyens pour le développement de l'ensemble du processus relative à ce pilier.
La conception et la réalisation de nos contributions seront présentées en détail dans le chapitre suivant du présent document.

Chapitre 3

Contribution au lanagement de la recherche scientifique

3.1 Introduction

L'objectif de ce chapitre est de présenter notre contribution au management de la recherche scientifique afin de répondre aux problématiques soulevées dans le chapitre précédent. Nous présentons, en premier lieu, notre démarche pour le développement d'une plateforme de management qui permet de gérer les nécessités du pilier de la recherche scientifique adaptés aux problèmes soulevés par les chercheurs et selon les objectifs fixés par les dirigeants.
Ainsi, nous allons étudier la problématique de l'adaptation des problèmes des chercheurs aux objectifs de développement d'un système d'information adéquat comme un problème d'optimisation en utilisant un PGI open source.
L'ambition d'un système d'information d'aide au pilotage de la recherche scientifique au sein d'un établissement ou d'une unité de recherche consiste à capitaliser les données récoltées et à les consolider de manière interactive en impliquant les différents acteurs de la communauté scientifique.
Ainsi, ce système d'information doit non seulement permettre de mutualiser les efforts de capitalisation des actions de recherche pour l'ensemble d'un établissement mais doit aussi répondre aux besoins propres de chaque acteur de la recherche scientifique relevant de cet établissement.

3.2 Les systèmes d'information

C'est un ensemble organisé de ressources : matériel, logiciel, personnel, données, et de procédures. Il permet d'acquérir, traiter, stocker, et de communiquer des informations (sous formes de données, textes, images, sons, etc.) dans des organisations. Toute organisation humaine (une entreprise, établissement ...) peut être perçue comme un système. Ce dernier peut être défini comme un ensemble d'éléments en interaction dynamique [33],

organisé en fonction d'un but donné. Pour parvenir à ce but, ce système tient compte de son environnement et régularise son fonctionnement en s'adaptant aux changements [34]. L'interaction entre ce système et son environnement est possible grâce à des flux d'informations. Ces flux circulent à l'intérieur, ce qui lui permet d'analyser son propre fonctionnement. Les éléments de ce système sont eux-mêmes des systèmes (ou sous-systèmes) : le système de décision exploite les informations qui circulent et organise le fonctionnement du système [35]. Des informations sont alors émises en direction du système opérant qui se charge de réaliser les tâches qui lui sont confiées. Il génère à son tour des informations en direction du système de décision qui peut ainsi contrôler les écarts et agir en conséquence.
Un SI a deux fonctions principales :

3.2.1 La production d'information

- Recueil de l'information : pour fonctionner, le système doit être alimenté. Les informations proviennent de différentes sources, internes ou externes. Les sources externes proviennent de l'environnement du système [36]. Il s'agit généralement de flux en provenance des partenaires du système (client, fournisseurs, administrations...). De plus en plus, l'établissement doit être à l'écoute de son environnement pour anticiper les changements et adapter son fonctionnement. Les développements des moyens de communication (internet en particulier) permettent de trouver facilement de l'information mais son exploitation reste délicate (qualité et fiabilité des informations). En interne, le système d'information doit être alimenté par les flux générés par les différents acteurs du système. Ces flux résultent de l'activité du système : approvisionnement, production, gestion des salariés, comptabilité, ventes ; la plupart de ces flux sont parfaitement formalisés (existence de procédures bien définies) mais, il existe également des flux d'information informelle (climat social, savoir-faire non formalisés,) qui sont par définition très difficiles à recueillir et à exploiter mais qui ont parfois beaucoup d'importance.
- Traiter et transmettre des informations : pour être exploitable, l'information subit des traitements. Là encore, les traitements peuvent être manuels (c'est de moins en moins souvent le cas) ou automatisés (réalisés par des ordinateurs).
- Mémoriser des informations : une fois l'information saisie, il faut en assurer la pérennité, c'est à dire garantir un stockage durable et fiable.
- Diffusion de l'information : pour être exploitée, l'information doit parvenir dans les meilleurs délais à son destinataire. Les moyens de diffusion de l'information sont multiples : support papier, forme orale et de plus en plus souvent, utilisation de supports numériques qui garantissent une vitesse de transmission optimale et la possibilité de toucher un maximum d'interlocuteurs.

3.2.2 La mise en œuvre d'outils de gestion

- Fonction technologiques (matériels, logiciels, méthodes, savoir-faire, ...) : les technologies mises en place, qu'il s'agisse des applications, des ordinateurs ou des réseaux fluidifient le cycle des informations aussi bien à l'intérieur qu'à l'extérieur de l'établissement. La question des outils est cruciale et la domination d'internet accélère l'innovation.
- Fonction économique.
- Fonction sociale : il faut noter que le (SI) a une autre finalité qui concerne la vie dans l'établissement, il doit permettre l'intégration.

3.3 Rôles du système d'information

Le système d'information représente le cœur de l'organisation interne, il gère l'information dans tous les niveaux et dans toutes les fonctions, cette information est celle qui représente le moyen primordial pour la prise de décision [37].
- Produire les informations réclamées par l'environnement.
- Déclencher les décisions programmées.
- Fournir des informations aux décideurs pour aider à la prise de décisions non programmées.
- Coordonner les tâches en assurant les communications au sein du système organisationnel.

3.3.1 Système d'information scientifique

Un système d'information scientifique [38] peut être défini comme un système informatique destiné à faciliter la gestion de l'ensemble des informations des chercheurs d'un établissement. Il s'agit d'améliorer la gestion des structures de recherche et des départements dans l'établissement.
Les SI scientifiques ont pour objectif d'améliorer le fonctionnement de tous établissements et faciliter l'inscription, le suivi des chercheurs, parmi ses objectifs nous pouvons citer la gestion des données qui sont utiles à l'ensemble des acteurs de la recherche scientifique comme par exemple :
- Travaux académiques non référencés dans les bases de données.
- Livres ou chapitres de livre publiés.
- Brevets déposés.
- Participation en tant qu'expert à des comités.
- Participation à des projets de recherche nationaux ou transnationaux.
- Encadrement de thèses.
- Contrats et partenariats avec des entreprises.

3.3.2 Les acteurs

Parmi les différents acteurs en interaction avec notre système nous trouvons des :
- Doctorants.
- Professeurs.
- Chefs d'établissement.
- Doyens.
- Présidents.
- Directeurs de thèse.
- Directeurs d'entité de recherche.

3.4 Progiciel Odoo

Le progiciel de gestion intégré est une catégorie de logiciel de gestion d'entreprise. C'est généralement une suite d'applications intégrées qu'une organisation peut utiliser pour collecter, stocker, gérer et interpréter les données de nombreuses activités commerciales.
Parmi les progiciels de gestion intégré open source nous nous intéressons à l'utilisation de Odoo [39] qui est une suite d'applications d'entreprise open source qui couvre tous les besoins. Sa valeur unique est d'être en même temps, très facile à utiliser et totalement intégrable.
Odoo est un progiciel open source de gestion intégré qui contient un grand nombre de modules qui permettent de simplifier la gestion d'entreprise en générale et plus précisément destiné à intégrer l'ensemble des données opérationnelles et de gestion de l'entreprise dans une base de données unique, accessible par une interface web.
Cette base de données centrale est associée à une couche fonctionnelle très innovante qui met en relation des informations d'origines diverses et qui assure un déroulement efficace des processus transversaux de création de valeur ajoutée de l'entreprise.
Le logiciel est utilisé par plus de deux millions d'utilisateurs pour gérer leurs établissements à travers le monde. Odoo est le système ERP open-source le plus populaire. Ce logiciel compte plus de 260 modules officiels et 7300 modules communautaires.

3.4.1 Architecture Odoo

MVC [40] est un modèle de conception qui décrit une architecture d'application informatique en la décomposant en 3 parties : modèle, vue et contrôleur.
Modèle d'architecture logicielle pour implémenter les interfaces utilisateur sur les ordinateurs. Il divise un logiciel donné en trois parties interconnectées de manière à faire des représentations internes distinctes des informations de la façon dont l'information est présentée ou reçue de l'utilisateur, nous trouvons :

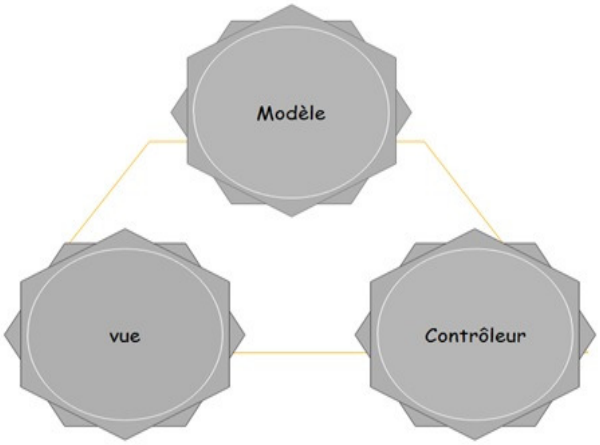

FIGURE 3.1 – *Architecture MVC.*

- Modèle : chaque objet déclaré dans Odoo correspond à un modèle, il est mappé à une table dans PostgreSQL.
- Vue : est l'ensemble des fichiers XML dans Odoo.
- Contrôleur : sont des classes Python qui gèrent la partie contrôleur.

3.4.2 Architecture d'un module

L'architecture modulaire d'Odoo lui permet de s'adapter à l'évolution des besoins dans le temps. Il s'agit de la faculté de construire des applications informatiques de manière modulaire (modules indépendants entre eux) tout en partageant une base de données unique, ceci élimine les saisies multiples et élimine l'ambiguïté des données de même nature.
Un module Odoo est caractérisé par les points suivants :
- Les vues, sous forme de fichiers XML. Ces vues sont sous forme de formulaires, listes, graphes, calendriers, ou de diagrammes.
- Les objets, sous forme de code python pour la plupart, ils contiennent les business objets et se chargent des traitements effectués par le module.
- Les workflows, sont des fichiers XML, permettant de modéliser les flux d'un état à l'autre.
- Les wizards, permettent l'affichage de fenêtres de dialogues, elles-mêmes contenant des vues ou des objets.
- Les rapports sont composés de fichiers XML pour la partie statique, de code python pour la partie dynamique et la mise en page se fait à l'aide d'OpenOffice.

3.4.3 Langages et technologies Utilisées

-Python [41] est un langage de programmation de haut niveau, orienté objet, totalement libre et efficace, conçu pour produire du code de qualité, portable et facile à intégrer. Ainsi la conception d'un programme Python est très rapide et offre au développeur une bonne productivité. En tant que langage dynamique, il est très souple d'utilisation et

constitue un complément idéal à des langages compilés.
- XML [42] est un langage informatique de balisage générique, qui permet de structurer des données afin quélles soient lisibles aussi bien par les humains que par des programmes de toute sorte. Il est souvent utilisé pour faire des échanges de données entre un programme et un serveur ou entre plusieurs programmes.
- Qweb [43]est le principal moteur de modélisation utilisé par Odoo. C'est un langage de template XML, Il s'agit d'un moteur de modèles XML utilisé principalement pour générer des fragments HTML et des pages. En utilisant le Qweb, on peut soit modifier les rapports déjà existants ou bien créer des nouveaux rapports.
- PostgreSQL [44] est un système de gestion de base de données relationnelles objet (ORDBMS) basé sur POSTGRES, version 4.2, développé à l'Université de Californie á Berkeley au département dínformatique. PostgreSQL a été le pionnier de nombreux concepts qui sont devenus disponibles dans certains systèmes de bases de données commerciales beaucoup plus tard.

3.5 Architecture de notre système de management de la recherche scientifique

La figure 3.2 illustre notre système conçu pour gérer le pilier de la recherche scientifique au sein d'une université publique, notre système cherche un parcours optimal à partir du problème des chercheurs en passant par des objectifs d'optimisation pour les dirigeants.

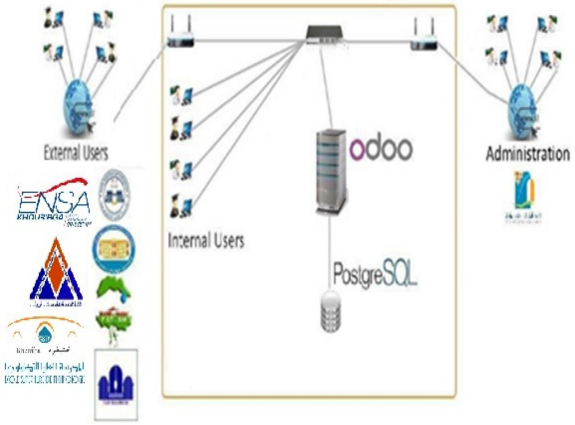

Figure 3.2 – *Architecture générale de notre système.*

Ainsi, nous nous focalisons sur l'adaptation du contenu au besoin, et plus précisément à la proposition des rubriques convenables au problème des chercheurs. Ces rubriques sont obtenues selon les besoins manifestés des chercheurs ainsi que le personnel administratif chargés de la gestion de la recherche scientifique. Une fois que la définition du

problème établie, nous passons à la personnalisation du contenu selon des droits d'accès stricts au responsabilité de chaque utilisateur.

3.6 La conception de notre système (Modélisation UML)

Dans notre système, nous proposons des rubriques qui facilitent la tâche des chercheurs par rapport à ces différents besoins vis-à-vis du système. Les tâches que notre système doit assurer sont :
- Affecter à chaque chercheur, un ensemble des rubriques adapté à son profil.
- Fournir au dirigeants les différents indicateurs concernant la recherche scientifique de l'université.
- Automatiser toutes les activités concernant la recherche scientifique.

3.6.1 Diagramme de cas d'utilisation

Le diagramme de cas d'utilisation représente la structure des grandes fonctionnalités nécessaires aux utilisateurs du système. Il donne une vue du système dans son environnement extérieur et définit la relation entre l'utilisateur et les éléments que le système met en œuvre. Les figures 3.3, 3.4 représentent respectivement les diagrammes de cas d'utilisation des chercheurs, et administrateur.

Figure 3.3 – *Diagramme de cas d'utilisation de l'acteur chercheur.*

3.6.2 Diagramme de classes

Le diagramme de classes est considéré comme le plus important de la modélisation orientée objet, il est le seul à être présent lors d'une telle modélisation. Alors que le diagramme de cas d'utilisation montre un système du point de vue des acteurs, le diagramme

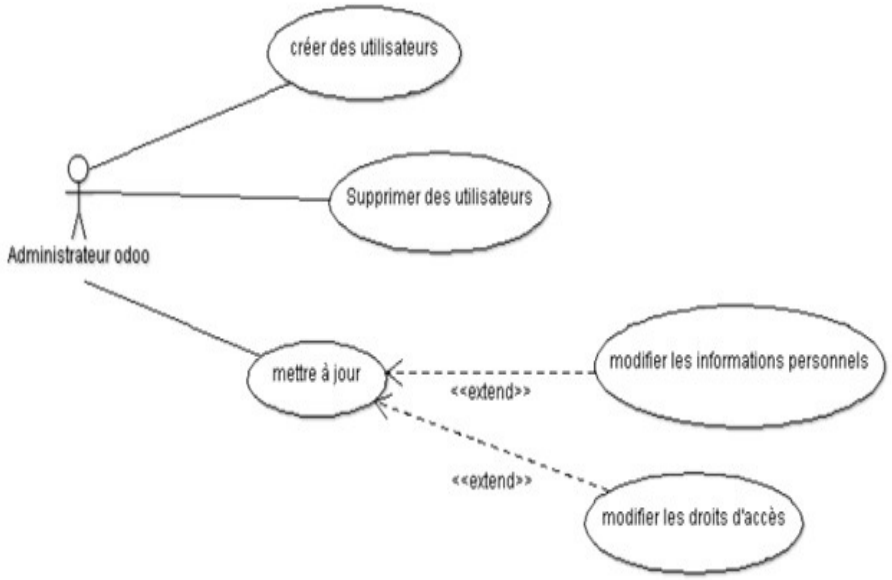

FIGURE 3.4 – *Diagramme de cas d'utilisation de l'administrateur.*

de classes en montre la structure interne. Ce diagramme permet de fournir une représentation abstraite des objets du système qui vont interagir ensemble pour réaliser les cas d'utilisation. La figure 3.5 représente le diagramme de classe de notre système.

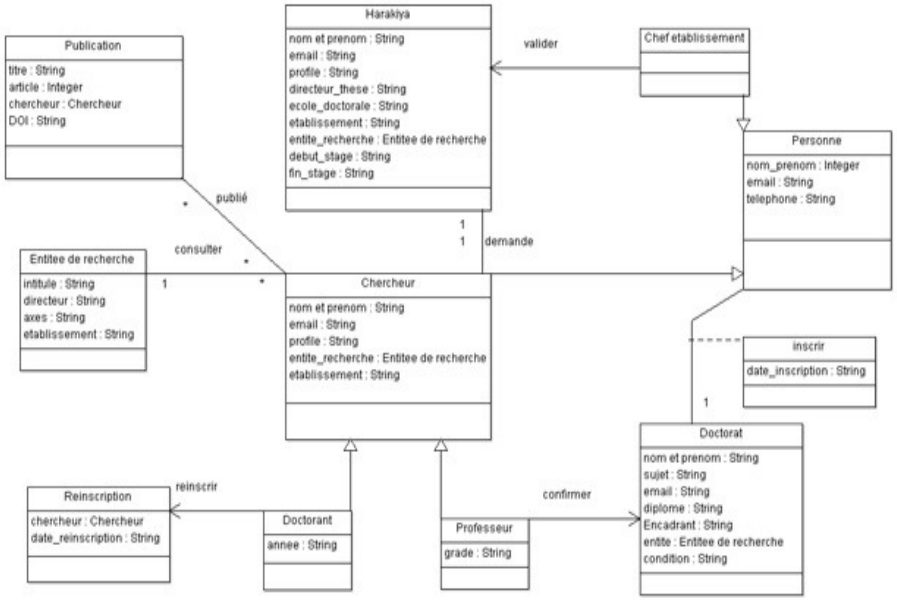

FIGURE 3.5 – *Diagramme de classe de notre système.*

3.6.3 Diagramme d'activité

Un diagramme d'activité est utilisé pour afficher la séquence des activités, il représente le flux de travail à partir d'un point de départ au point d'arrivée, détaillant les nombreux sentiers de décision, qui existent dans la progression des événements contenus dans l'activité.

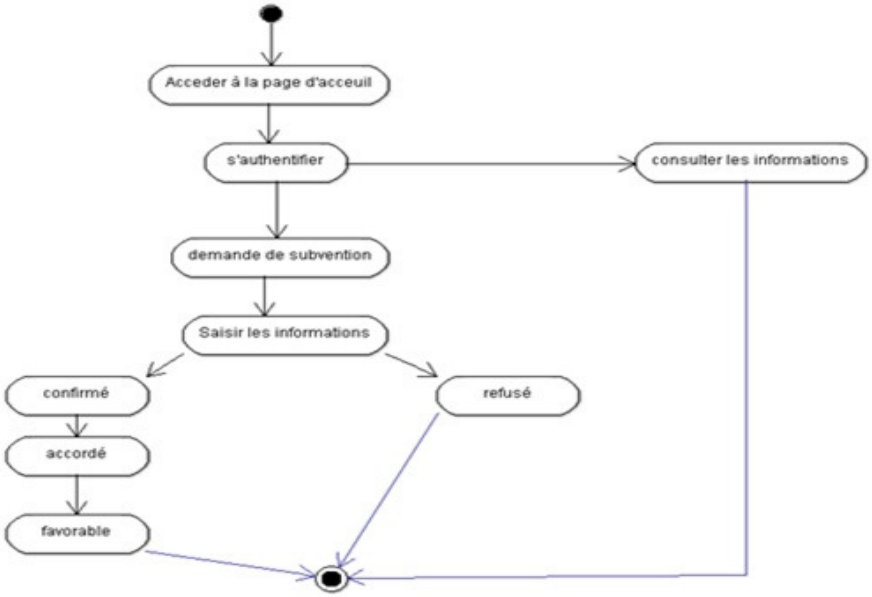

FIGURE 3.6 – *Diagramme d'activité.*

3.7 IMPLÉMENTATION

Le système a été conçue pour faciliter son utilisation et réduit les couts associés, pour comprendre l'intérêt de notre système, imaginons un étudiant intéressé par la pré-inscription à un sujet de thèse sur notre système. Après son inscription via le module portal (ce module est dédié au utilisateurs externe du système), l'étudiant doit remplir un formulaire en ligne. Pour définir son sujet de pré-candidature. Ensuite, notre système fournit à cet étudiant un moyen de suivi de sa candidature jusqu'à une réponse finale.
Après l'authentification de l'enseignant qui a déposé le sujet de la thèse, la fenêtre illustrée sur la figure 3.7 s'affiche. Ce formulaire permet à l'enseignant d'évaluer les candidatures reçues pour un sujet de thèse donnée.

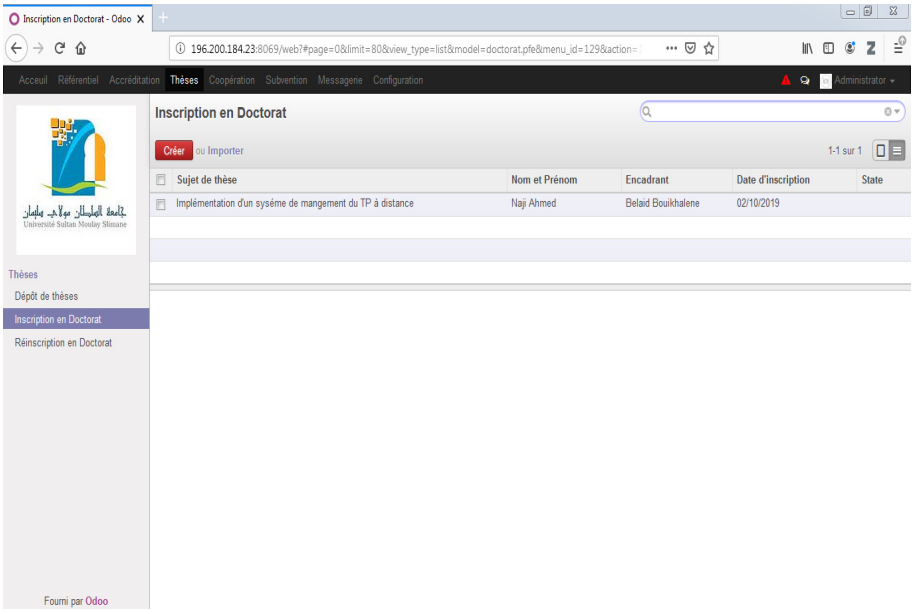

FIGURE 3.7 – *Formulaire des inscriptions sur un sujet de thèse.*

La fenêtre illustrée sur la figure 3.8 permet à l'enseignant d'ajouter un nouveau sujet de thèse en respectant les lois en vigueur (ne pas encadrer plus que cinq doctorants à la fois, la description de sujet de thèse, les prérequis et d'autres informations.)

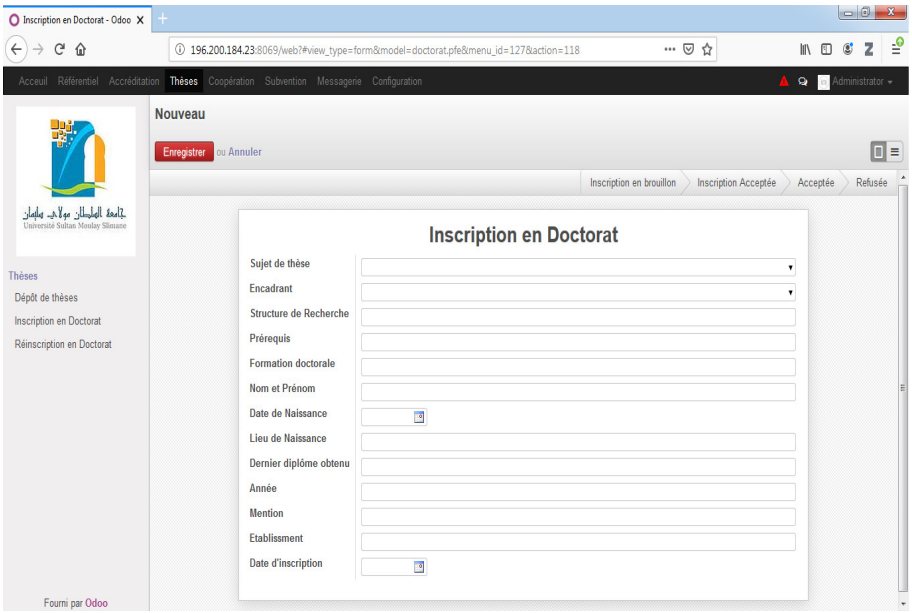

FIGURE 3.8 – *Formulaire de dépôt du sujet de thèse.*

La fenêtre illustrée sur la figure 3.9 permet au doctorant de procéder à la réinscription à travers notre système en fournissant un rapport d'activité des actes scientifiques réalisés, etc.

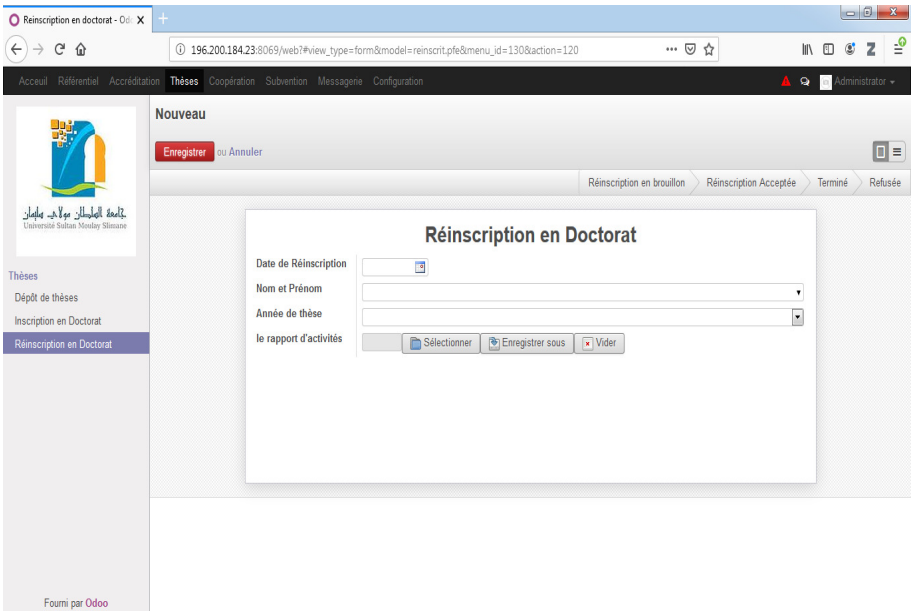

FIGURE 3.9 – *Formulaire de réinscription en doctorat.*

Par ailleurs, pour soutenir une thèse comme nous illustre la figure 3.10, le doctorant doit avoir comme connaissances préliminaires certains conditions obligatoires (Troisième inscription administrative, publications des papiers scientifiques dans une base de données indexée, et d'autres conditions.)

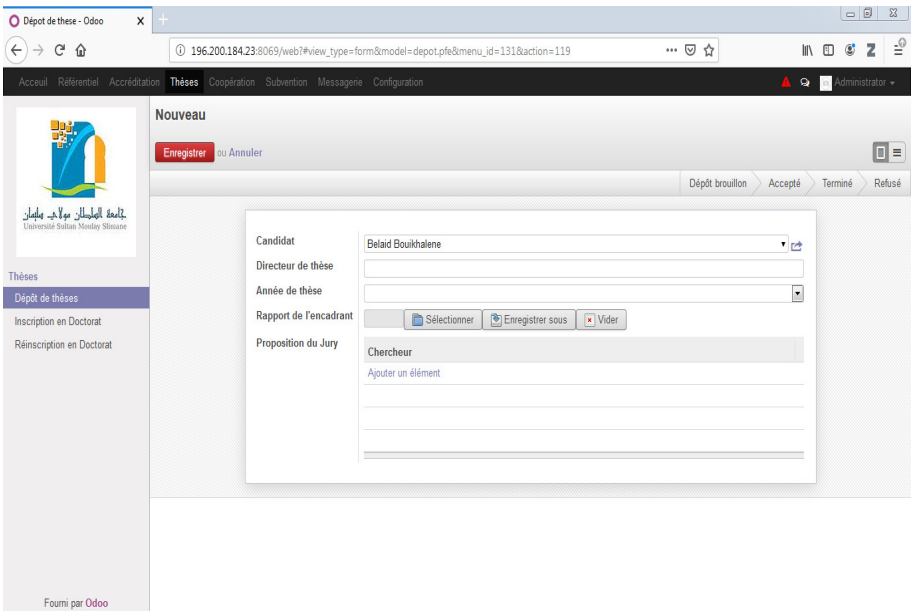

FIGURE 3.10 – *Formulaire de dépôt de la thèse du doctorat pour la soutenance.*

La fenêtre illustrée sur la figure 3.11 permet aux dirigeants d'avoir une vue en temps réelle sur l'ensemble des actions de recherche (nombre de publication par établissement,

nombre de thèse soutenu, nombre d'évènement organisé par chaque établissement ou structure de recherche,...).

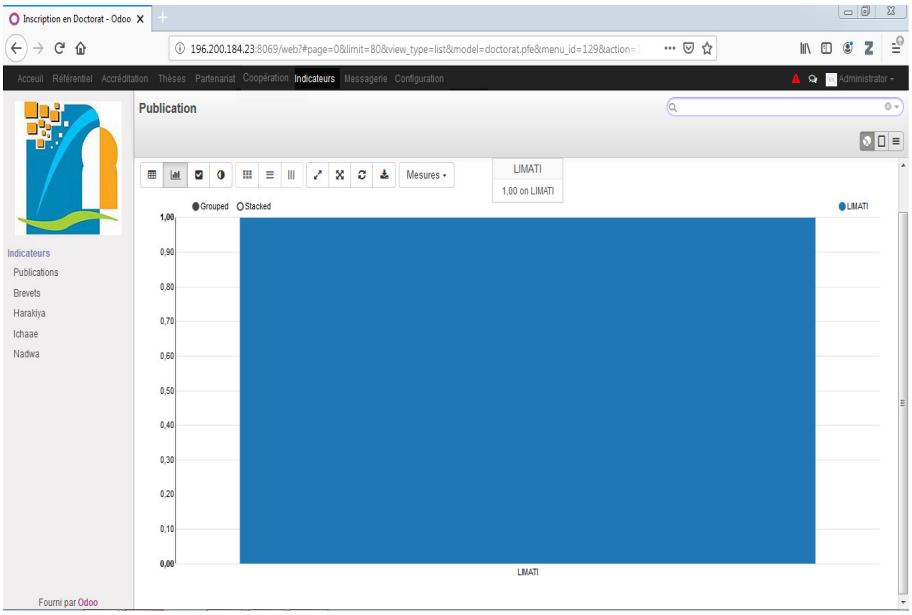

FIGURE 3.11 – *Interface des différents indicateurs.*

3.8 CONCLUSION

Nous avons présenté Odoo pour comprendre leurs fonctionnements et leurs architectures afin d'avoir une idée précise sur l'utilité et la nécessité d'utilisation des PGI.
A savoir que dans la réalisation de ce système, on a eu recours à un ensemble de technologies, comme python qui est le langage de programmation utilisé dans le développement, comme PostgreSQL la gestion des bases données, et XML et CSS pour le côté vue et design de l'applications.
Dans ce chapitre, nous avons présenté la première partie de nos contributions ; premièrement, nous avons présenté un système d'information scientifique adaptée à l'université. L'idée est de résoudre le problème du management de la recherche scientifique à travers l'automatisation de l'ensemble des actions liée à se pilier primordial pour l'université.
Par ailleurs, nos contributions ne se limitent pas à l'automatisation de la recherche scientifique. Outre cela, dans le chapitre suivant, nous allons soulever la problématique d'automatisation du traitement des CV des chercheurs. C'est pourquoi, nous proposons une approche de la prédiction de leurs domaines du recherche en utilisant des algorithmes d'exploration de données.

Chapitre 4

Vers un traitement automatique des curriculum vitæ des chercheurs

4.1 Introduction

La problématique élaborée dans ce chapitre se situe dans le cadre général des méthodes de traitement automatique. En effet, le traitement automatique des documents[45] est un domaine de recherche prometteur, car elle garantit un accès rapide et ciblé à l'information.

Le problème de traitement des documents [46] se présente pour les établissements qui possèdent des milliers de documents dans leur archive dans des bases de données sans pouvoir accéder à l'information dont elles ont besoin. Pour faciliter la recherche documentaire [47], ces établissements font souvent appel à des techniques déjà disponibles avec les moteurs de recherche qui utilisent en principe la correspondance [48]. Ce qui ne garantit pas des solutions pertinentes qui peuvent réduire la complexité des données pour tirer seulement l'information utile.

Il est très fréquent qu'une recherche ne retourne pas les documents dont nous avons besoin. D'ailleurs, même si les documents retournés contiennent ce qui était recherché, l'utilisateur doit par la suite fouiller le texte en question pour trouver l'information dont il a besoin, ce qui pose problème dans l'effort de l'automatisation du traitement automatique des données [49].

Dans cette perspective, nous nous intéressons particulièrement à la description du contenu des curriculum vitæ des chercheurs par des descripteurs sémantiques [50], plus précisément, l'indexation sémantique d'un corpus [51] (collection de CV des chercheurs). En effet, la recommandation sémantique d'un document s'effectue par le calcul de la similarité sémantique entre un document [52] et chaque document du corpus.

En somme, les développements réalisés dans le cadre de ce chapitre se composent de trois axes principaux. Nous présentons d'abord, les systèmes de recommandation [53] en particulier la recommandation basée sur le contenu [54]. Ensuite, nous introduisons la notion de l'indexation sémantique et le calcul de la similarité sémantique en se basant sur un

corpus textuel. Enfin, nous terminons par la prédiction des domaines de recherches des chercheurs en exploitant des algorithmes d'apprentissage supervisés.

4.2 Le traitement automatique de la langue

Le traitement du langage naturel (TALN) [55] imite la compréhension humaine des mots et des phrases et permet désormais aux modèles d'apprentissage automatique [56] de traiter d'énorme quantités d'informations avant de fournir des réponses précises aux questions qui leur sont posées.
Dans une vision d'automatisation, l'abondance des quantités d'informations disponibles qui dépassent les capacités d'assimilation, nécessite l'adoption des méthodes qui sont capables de tirer l'information à partir de toutes ces masses de données qui deviennent une nécessité.
Cette discipline récente est née aux États-Unis vers 1942, le traitement du langage naturel construit un carrefour pour plusieurs autres approches comme l'approche linguistique [57] qui travaille la phrase sur sa totalité comme objet de traitement, aussi l'approche syntaxique basé sur la théorie des langages nécessite des efforts pour l'analyse des phrases puis leur intégration en un ensemble adéquat ce qui est pas le cas en générale, et l'approche numérique plutôt basé sur les fondamentales mathématiques de la probabilité et statistique qui vise à calculer les probabilités de cooccurrences entre les mots ou les expressions dans une masse importante de données. C'est ce qui est demandées à nos jour et qui est non relatif à une langue donnée ce qui lui rend plus crédible parmi les différentes approches. Parmi les champs d'applications de ces approches, nous citons :
- La syntaxique [58] : séparation des mots, délimitation de phrase, etc.
- La sémantique [59] : traduction automatique, détection de coréférence et la résolution d'anaphore, etc.
- Le traitement du signal [60] : traitement de la parole, détection des langues et des dialectes, etc.
- L'extraction d'informations [61] : analyse des sentiments, recherche d'informations, classification et catégorisation de documents, etc.
- La bibliométrie [62] : traitement des publications scientifiques, etc.

Par ailleurs, dans le cadre de cette thèse, nous nous sommes intéressés à la problématique du traitement automatique du curriculum vitæ par une contribution intitulée traitement automatique de CV pour la recherche scientifique utilisant un algorithme d'exploration de données [63]. Le détail de cette contribution sera présenté profondément dans le chapitre 4.
Nous proposons dans la suite les principaux fondements de notre travail et le traitement automatique du curriculum vitæ des chercheurs. L'idée de cette approche est le calcul de la similarité sémantique entre les documents d'un corpus textuel ensuite l'application des algorithmes d'apprentissage supervisée.

4.3 Le traitement automatique de la langue

L'évolution des nouvelles technologies de l'information et de la communication est tout ce qui concerne l'archivage et le stockage des documents numériques était accompagné par le développement des méthodes pour extraire des informations pertinentes de cette masse d'information. Pour que ces méthodes seront efficaces, nous aurons besoins d'extraire des descripteurs qui seront automatisés et crédibles, l'extraction des descripteurs construit un champ de recherche en plein expansion.
L'indexation du document [64] constitue un processus pour analyser et spécifier un ensemble des descriptives significatives d'un document nommé «les descripteurs», Le processus de représentation de ces documents est appelé le processus d'indexation.
Le traitement automatique du langage naturel [146] propose des méthodes qui essaye d'automatiser l'extraction des descripteurs ; en revanche, les indexations manuelles qui nécessite l'implication d'un documentaliste qui essaye d'établir une liste des descripteurs d'une façon manuelle. Les résultats obtenus par une indexation automatique sont souvent jugés insatisfaisants. Pour remplir ces lacunes, certaines recherches proposent de valider cette liste des descripteurs par des documentalistes spécialisés dans un domaine donné, ce qui nous amène à une indexation semi-automatique ou indexation supervisée.
Dans les différents processus d'indexation manuelle, supervisée ou semi-supervisée, la définition d'un ensemble de descripteurs qui représente le langage d'indexation est associé à chaque document du corpus. Cet ensemble des descripteurs nécessite une validation par un documentaliste spécialiste dans son domaine.
Chaque type d'indexation peut être contrôlé par un expert dans le domaine comme celle du descripteur manuel ou semi-automatique, contrairement au processus d'indexation automatique qui propose tous les descripteurs issus de l'analyse automatique des documents du corpus son validation par un expert.
Pour chaque corpus donné qui construit des documents et qui touche le même axe de données dans le domaine scientifique, littérature ou linguistique, ...Ces corpus construisent une mine précieuse de données pour tous les traitements automatiques du langage naturel qui seront validés par des experts. Ces corpus permettent de dégager des données précieuses pour les approches numériques qui sont basées sur le mathématique et spécialement en statistique et probabilité, afin de valider ou réfuter les hypothèses enlevées.
Pour les experts, deux méthodes d'évaluations peuvent être abordées [66], d'une part les méthodes intrinsèques qui se basent sur la vérification manuelle des résultats produits soit d'une façon automatique en calculant les mesures de similarités ; d'autres part, nous trouvons les méthodes extrinsèques qui essayent d'évaluer sans juger en accélérant l'automatisation des tâches.

4.4 Processus d'extraction des termes pertinents

Dans cette section, nous présentons la démarche suivie pour extraire automatiquement les descripteurs à partir d'un corpus donné. Les étapes de cette démarche sont :

4.4.1 Le modèle vectoriel

Nous avons choisi d'appliquer des méthodes numériques [67] pour tirer parti de leur efficacité. Pour appliquer ces techniques, les textes doivent être convertis en représentation pour permettre d'appliquer des calculs.
Dans cette section, nous expliquerons brièvement comment utiliser ces méthodes pour transformer des textes en vecteurs ? et aussi comment traiter les problèmes inhérents à de telles représentations ?
La première étape vers l'application du modèle vectoriel [68] sur un corpus est le choix d'unités textuelles ou de termes d'indexation (termes, n-grammes de termes ou caractères, expressions). Ces termes vont constituer notre vocabulaire utilisé.
Chaque élément du vocabulaire est associé à un index unique. Ensuite, un vecteur V est attribué à chaque segment de texte (une phrase, un paragraphe, un document). La dimension de ce vecteur correspond à la taille du vocabulaire et chaque composante V_i associe une pondération au terme d'indice i (par exemple la fréquence d'apparition du terme dans le segment).
Dans cet espace, chaque document est représenté par un vecteur calculé à partir des unités textuelles les plus pertinentes de chaque segment. Ainsi, il est souhaitable, de comparer les paragraphes des phrases qu'ils contiennent ou les phrases de leurs termes ou de comparer les paragraphes avec les termes de leurs personnages.
Pour illustrer cette approche, nous prenons le document du tableau 4.1.

2008-2009 :	DUT in Computer Engineering.
2014-2015 :	Master's Degree Computer Science and Systems.

Table 4.1 – *Extrait du CV.*

Nous choisissons comme vocabulaire les termes séparés par des espaces. Dans la représentation vectorielle, chacune des deux phrases sera un vecteur dont les composantes indiquent la fréquence (1) ou l'absence (0) d'un terme dans la phrase (voir tableau 4.2).

Sentence	2008	-	2009	:	DUT	In	Computer	Engineering	.	2014	2015	Master	Degree	Science	And	Systems
1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
2	0	1	0	1	0	0	1	0	1	1	1	1	1	1	1	1

Table 4.2 – *Matrice des segments de termes pour le texte du tableau.*

La disposition des vecteurs forme la matrice des termes et des phrases. La variabilité et la complexité des informations textuelles, ainsi que la taille importante des collections

de documents qui ne permettent pas aux modèles complexes de traiter rapidement et automatiquement des données importantes et la représentation sophistiquée des documents. Cela nous donne une exploitation statistique des données, mais ne permet pas de comprendre le sens du texte.
Pour faire des calculs de fréquence ou des distributions, il est très utile d'avoir une correspondance entre les termes et les composantes des vecteurs des phrases.
Cet exemple utilise jusqu'ici les poids tf (fréquence des termes) [69]. La mesure du nombre d'occurrence d'un terme dans la collection (tf) ne rend pas compte de sa spécificité. Cependant, un terme commun à trop de documents est moins utile qu'un autre terme commun à quelques-uns seulement. Nous allons utiliser tf-idf [70] qui est une combinaison du nombre d'occurrence du terme dans le document et de la valeur inverse du nombre de documents dans lesquels il est présent idf (fréquence de document inversée) [71] , définie dans :

$$Tf.idf_i = \frac{n_{i,j}}{\Sigma_j n_j} \times \frac{\log |D|}{|\{dj : ti \in dj\}|} \tag{4.1}$$

avec | D | : Nombre total de documents dans le corpus.
ni, j : Nombre d'occurrence de terme ti dans le document dj.
nj : Nombre d'occurrence de tous les mots du document.
$|\{dj : ti \in dj\}|$:Nombre des documents ou le terme ti apparait au moins une fois.

4.4.2 Réduction dimensionnelle : prétraitement linguistique

Pour définir les unités textuelles d'un corpus, nous utilisons les "mots" pouvant être produits par de simples techniques de segmentation automatique [72]. Cependant, ces unités de base peuvent également être traitées pour intégrer les connaissances linguistiques aux représentations.
Ensuite, le lexique joue un rôle important dans la composition de la matrice, car nous utilisons différents processus pour réduire la malédiction de la dimensionnalité. Les méthodes suivantes sont appliquées :
- Filtration : il est inutile d'indexer ou d'utiliser certain mot dans un processus de recherche d'information, nous devons procéder à la suppression de verbes et des mots fonctionnels (être, avoir, pouvoir, avoir, être ...), ainsi des expressions communes (par exemple, chacune de ...), des chiffres ou des lettres) et des symboles (tels que $, #,etc.).
- Détection de mots composés : le problème des mots composés est automatiquement marqué puis transformé en un terme enraciné unique.
- Harmonisation de la casse : transformer tout le texte du document en minuscule pour unifier le traitement.
- Racinisation : le processus de supprimer le suffixe et le préfixe d'un mot en laissant son radical.
- La lemmatisation [73] : en trouvant la racine des verbes repliés et en renvoyant les mots

pluriels féminins ou masculin au singulier avant de les associer à un certain nombre d'occurrences, en regroupant les différentes formes que peut revêtir un mot.

Sentence 1	dut in computer engineering
Sentence 2	master degree in computer science and systems

TABLE 4.3 – *Texte du tableau 4.1 après prétraitement linguistique.*

Ces opérations réduisent considérablement la taille de l'espace tout en augmentant la fréquence des termes canoniques, malgré une perte d'informations, ce qui est normal après cette ensemble d'opérations.

Sentence	dut	in	computer	engineering	master	degree	science	systems
1	1	1	1	1	0	0	0	0
2	0	0	1	0	1	1	1	1

TABLE 4.4 – *Matrices réduites des segments des termes après le prétraitement.*

4.4.3 La similitude de vecteur

À partir d'une représentation mathématique, une méthode de calcul de la proximité entre les unités textuelles consiste à utiliser des mesures communes de la similarité vectorielle [74], comme par exemple le cosinus :

$$\cos(D_a, D_b) = \frac{|D_a.D_b|}{\| D_a \| \| D_b \|} \tag{4.2}$$

avec :

D_a : Le vecteur obtenu pour les documents a.
D_b : le vecteur obtenu pour le document b.

Plus l'angle entre eux est petit, plus les informations qu'ils transportent sont proches, la figure 4.1 montre un exemple avec 2 documents, A et B. Ainsi l'angle θ est nulle et montre que les documents A et B n'auront aucun mot en commun relativement proches, tandis que, pour les autres cas, nous pouvons dire qu'il existe une sorte de similarité qui dépend du résultat du cosinus et qui varie entre les l'intervalle [-1,1].

D'autres méthodes de mesure peuvent être utilisées en analyse textuelle, nous trouvons la similarité vectorielle. Par exemple, en utilisant la distance du khi-deux [75] qui présente des résultats précis et qui ne nécessitent aucune interprétation supplémentaire, le cosinus et le Kullback-Leibler [76] mesurent un corpus textuel.

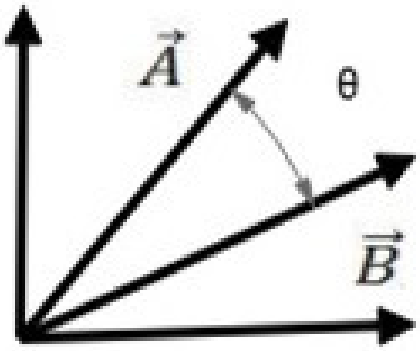

FIGURE 4.1 – *Représentation de proximité des documents par l'angle θ .*

4.5 La prédiction du domaine de la recherche pour le chercheur

Dans cette section, nous introduisons quelques algorithmes d'apprentissage automatique à savoir l'arbre de décision, naïve bayésienne et one rule. Ces algorithmes sont tous des algorithmes supervisés qui visent à apprendre et à améliorer les performances à travers les expériences. En revanche, avec l'apprentissage non supervisé ou l'apprentissage semi supervisé.
La capacité de prédire le domaine de la recherche des chercheurs nous redirige dans le champ d'étude de l'intelligence artificielle par le biais des algorithmes d'apprentissage supervisés. En effet, l'apprentissage à travers les arbres de décision, naïve bayésienne et one rule utilise des algorithmes sophistiqués et efficaces basés sur des modèles prédictifs [77]. Ces derniers constituent un outil d'aide à la décision pour l'évaluation de la valeur d'une caractéristique d'une population en se basant sur l'observation des autres caractéristiques de la même population.
Nous introduisons une description des algorithmes prédictif utilisé :

4.5.1 Arbre de décision

« Arbre de décision » [78] est un classifieur simple et largement utilisé. L'idée sous-jacente est d'appliquer une méthode simple pour résoudre le problème de classification [79]. Parmi les raisons pour laquelle nous avons choisis l'arbre de décision c'est qu'il fournit correctement la classe de l'échantillon d'entraînement le plus possible ; il se généralise également au-delà de l'échantillon d'apprentissage, de sorte que les échantillons invisibles puissent être classés avec la plus grande précision possible. L'arbre de décision est facile à mettre à jour car de plus en plus d'échantillons d'apprentissage sont disponibles et ont une structure simple à adapter.

4.5.2 Naïve bayésienne

« Naïve Bayes » [80] est un algorithme puissant et simple pour la classification, la probabilité bayésienne est l'estimation d'un événement. L'idée derrière ce type de classifieur est de prédire les probabilités d'appartenance à une classe, telles que la probabilité que les données appartiennent à une classe particulière. Les raisons de choisir Naïve Bayes, c'est que chaque hypothèse est associée à une probabilité. L'observation d'une ou plusieurs instances peut modifier cette probabilité. En outre, nous pouvons parler des hypothèses les plus probables, sur la base des probabilités conditionnelles et de la règle de Bayes.

4.5.3 One Rule

« One Rule » consiste à générer une règle pour chaque prédicteur dans les données, puis à sélectionner la règle avec les erreurs les plus faibles [81]. Il s'agit de la règle unique. Le choix de One Rule s'explique par le fait qu'une règle est créée pour chaque attribut et chaque valeur de cet attribut. De plus, il consiste à compter le nombre de fois que chaque classe apparaît et à trouver la classe la plus fréquente en calculant le taux d'erreur des règles, pour choisir les règles avec le taux d'erreur le plus faible.

4.6 Descriptif du modèle prédictif

Les chercheurs fournissent ces curriculums vitæ en remplissant ces informations générales à travers le système de management de recherche scientifique de l'université. Dans cette direction, les données sont extraites de la base de données Postgresql qui contient tous les Curriculum vitæ des chercheurs de l'université dans les différents domaines.

Figure 4.2 – *Architecture spécifique du modèle prédictif.*

Le modèle du figure 4.2, nous propose une idée générale sur le processus suivi dans cette

démarche afin de prédire le domaine de la recherche des chercheurs.
Notre but est de prédire le domaine de la recherche du chercheur à partir de son curriculum vitæ en se basant sur les données extraites par le biais des différents prétraitements déjà citées, la figure 4.3 présente les différents attributs utilisés dans notre étude :

```
@relation 'SRSMSU CV'

@attribute Text string
@attribute class {computer_science,physics,natural_science,geography,mathematics}

@data

'elmohadab mohamed\nn20 lot el arsa quartier Chara, ouled hamden beni mellal\nmaster_compu
'ouatik fahd\nn 17 taqadoum beni mellal\nmaster_computer_science_systems\nfahd.ouatike@gma
'abouhilal abdlmoula \nn 117 quartier moderne beni mellal\nmaster_computer_science_systems
'elhajri elmahdi \nn20 quartier choroque safi\nmaster_computer_science_systems\nelmahdi.el
```

FIGURE 4.3 – *Extrait des données.*

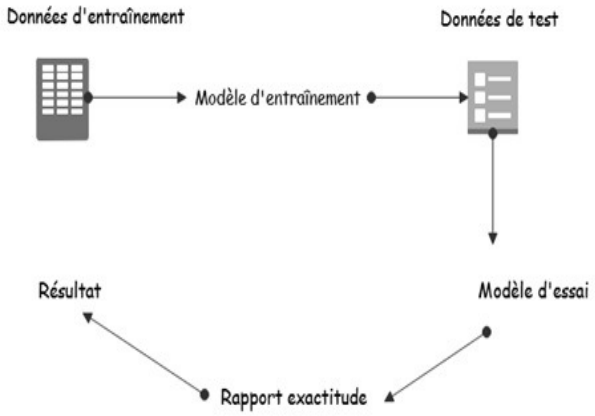

FIGURE 4.4 – *Processus généraux de la prédiction.*

La figure 4.4, nous fournit une idée globale sur le fonctionnement des processus qui débutent par l'acquisition des données prétraitées, après nous appliquons l'algorithme donné dans le but de construire un modèle de test après nous appliquons la prédiction sur ce dernier, nous comparons les différentes paramètre d'évaluation de chaque algorithme pour en tirer le plus adéquat dans notre cas.
Après le prétraitement déjà mentionné nous appliquons le filtre Sting to Word Vector afin d'obtenir l'ensemble des attributs sur lesquelles nous allons travailler [82]. D'abord, nous commençons par l'application des trois algorithmes sur nos données.

4.7 Evaluation d'une méthode de prédiction

Nous allons étudier dans cette partie comment nous pouvons évaluer la performance d'une méthode de prédiction pour s'assurer de sa capacité à satisfaire nos besoins. Le choix d'une mesure, doit être dépendant du type de données à traiter et nos intérêts souhaités. En partant du principe que la prédiction est une filiale de la recherche d'information [83]. L'évaluation de la performance d'une méthode de prédiction est manifesté à

travers plusieurs mesures [84].

Les paramètres d'analyse utilisés dans cette étude sont présentés de la manière suivante :
- Écart quadratique moyen : est une mesure caractérisant la précision de cet estimateur. Elle est plus souvent appelée « erreur quadratique », il est défini par :

$$\sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{n}} \tag{4.3}$$

- L'erreur absolue moyenne : est une mesure de la différence entre deux variables continues, il est défini par :

$$\frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{n} \tag{4.4}$$

- L'erreur absolue relative : est également relative à la moyenne des valeurs réelles. Dans ce cas, nous prenons l'erreur absolue totale et on la normalise en la divisant par l'erreur absolue totale du prédicteur simple, il est défini par :

$$\frac{|p_1 - a_1| + \cdots + |p_n - a_n|}{|a_1 - \bar{a}| + \cdots + |a_n - \bar{a}|} \tag{4.5}$$

- L'erreur relative au carré : est simplement la moyenne des valeurs réelles. Ainsi, il prend l'erreur au carré totale et la normalise en la divisant par l'erreur au carré totale du prédicteur simple, il est défini par :

$$\sqrt{\frac{(p_1 - a_1)^2 + \cdots + (p_n - a_n)^2}{(a_1 - \bar{a})^2 + \cdots + (a_n - \bar{a})^2}} \tag{4.6}$$

- Précision : également appelée la valeur prédictive positive, il s'agit de la fraction d'instance extraite qui est pertinente.

$$\text{Prćision} = \frac{N_{ps}}{N_p} \tag{4.7}$$

- Rappel : la sensibilité est la fraction d'instances pertinentes récupérées.

$$Rappel = \frac{N_{ps}}{N_p} \tag{4.8}$$

- F-Mesure : une mesure qui combine précision et rappel, c'est le moyen harmonique de précision et de rappel.

$$F - mesure = 2.\frac{(\text{Prćision}.Rappel)}{(\text{Prćision} + Rappel)} \tag{4.9}$$

Après l'application des trois classifieurs dans les données présentées à la figure 4.3, nous avons obtenu les résultats présentés au tableau suivant :

	Decision Tree	Naive Bayes	OneR
CCI	97.1042%	92.13%	95.67%
ICI	2.8958%	1.87%	4.33%
Kapp Statistic	0.643	0.562	0.612
Root Mean Squad Error	0.1077	0.113	0.1085
Relative Absolute Error	100%	95.9881%	98.861%
Root Relative Squad Error	100%	97.376%	98.346%

TABLE 4.5 – *Comparaison des performances 1.*

Le tableau 4.5 montre les résultats qui permettent de définir l'algorithme avec une bonne prédiction. Selon les résultats présentés, l'algorithme avec une CCI élevée est l'arbre de décision (CCI = 97.1024). En outre, pour obtenir des meilleurs résultats, l'erreur doit être minimale, c'est le cas pour l'arbre de décision car c'est le seul classifieur qui répond à cette condition. La statistique Kappa est un autre paramètre à prendre en considération. Si celui-ci est plus élevé, cela signifie donc une bonne qualité de prédiction. Comme le montre le tableau 4.5, l'arbre de décision est le classifieur dont la statistique Kappa est supérieure à (0. 643).

	Decision Tree	Naive Bayes	OneR
Precision	0.943	0.86	0.921
Recall	0.971	0.89	0.95
F-measure	0.957	0.875	0.932

TABLE 4.6 – *Comparaison des performances 2.*

Comme le montre le tableau 4.6, les résultats numériques se rapprochent les unes des autres. Les mesures de précision, de rappel et F-mesure doivent être proches de 1. Ainsi, une observation approfondie montre que l'arbre de décision est l'algorithme qui satisfait cette condition (précision = 0,943, rappel = 0,971 et F-mesures = 0,957).
En résumé, les trois algorithmes proposés ont donné de bons résultats, mais l'arbre de décision reste le meilleur en termes du résultat fournit par les paramètres d'évaluation.
Maintenant, nous passons à la phase de prédiction [85], nos données pour la prédiction sont :

```
@relation 'SRSMSU CV'

@attribute Text string
@attribute class {computer_science,physics,natural_science,geography,mathematics}

@data

'fadil mohamed elbachir\nn 89 lot quartier hamdania,beni mellal\ntelecommunications_en
'houkmi nabil\nn 99 quartier almostaqbal beni mellal\nenvironmental_engineer\nnabil.ho
'addaoui adam \nn 56 quartier ryadsallam beni mellal\nmaster_computer_science_systems\
'eltobi ali \nn 20 quartier nassim sale\nmaster_computer_science_systems\neltobi.ali@g
'elabdelaoui hassan \nn 78 quartier nassre meknes\nelectrical_engineer\nhassan.elabdel
'zyad amine \nn 16 quartier charaf beni mellal\nmaster_computer_science_systems\namine
'hanani ahmed \nN°13 zouair ouled yaich beni mellal\nelectrical_engineer\nhananine.ahm
'elhamri abdellatif \nn 36 qaurtie elouarda beni mellal\nmaster_mathematics\nabd.elham
'amrani mohamed \nn 3 qaurtie rida marrakech\nbiological_engineer\nmed.amrani@gmail.co
'naciri fadoua \nn 167 qaurtie moderne beni mellal\nmaster_mathematics\nfadoua.naciri@
```

FIGURE 4.5 – *Extrait des données pour la prédiction.*

Afin de prévoir une nouvelle classe pour le nouveau curriculum vitæ à partir de la

base des curriculum vitæ déjà existante dans la base de données, en combinant le filtre chaˆen vecteur word avec l'algorithme arbre de décision [86].

Instances	Predicted
1	Physics
2	Natural Science
3	Computer Science
4	Computer Science
5	Physics
6	Computer Science
7	Physics
8	Mathematics
9	Natural Science
10	Mathematics

TABLE 4.7 – *Extrait du résultat des 10 premier données de prédiction.*

4.8 CONCLUSION

Dans ce chapitre, nous avons abordé le traitement automatique de la langue avec ces différentes étapes du prétraitement linguistique comme filtration, lemmatisation, harmonisation de la casse, racinisation. Ensuite, nous avons présenté le processus d'extraction des termes pertinents en nous basant sur la similitude de vecteur, après nous avons décrit les différentes démarches suivies pendant le processus de prédiction.
Nous avons présenté également des algorithmes d'apprentissage supervisée utilisée dans ce travail. Après, nous avons cité, notre modèle prédictif. Finalement, nous avons introduit des mesures d'évaluations de la performance des méthodes de prédiction, dont la F-Mesures, la précision et le rappel sont les plus populaires.
Dans le chapitre suivant, nous allons soulever la problématique du classement du papier de recherche scientifique. Ainsi, nous proposons une approche de la prédiction du futur classement d'un papier scientifique dans leur domaine du recherche en utilisant des algorithmes d'apprentissage supervisés.

Chapitre 5

Contribution au classement des papiers scientifiques

5.1 Introduction

L'objectif de ce chapitre est de présenter notre contribution au classement des papiers scientifiques afin de répondre à une des problématiques soulevées dans les chapitres précédents. Nous présentons, en premier lieu notre démarche pour le développement d'une méthode de classement des papiers scientifiques qui permet de répondre aux quelques critiques adressées aux méthodes déjà existantes. Ainsi, nous proposons une méthode qui permet de prédire le classement des futurs papiers de recherches scientifiques publiés par les chercheurs, en utilisant les algorithmes d'apprentissage supervisés.
Dans ce qui suit, nous présentons d'abord un bref historique des méthodes du classement du papier scientifique déjà existant. Ensuite, nous enchaînons avec les critiques adressés aux méthodes de classement existant, puis nous discutons l'utilisation de quelques classifieurs d'apprentissage supervisé pour prédire le futur classement des papiers de recherches scientifiques.

5.2 Processus et modèles de classement

Le classement le plus adéquat pour des documents dépend de plusieurs paramètres qui sont au préalable demandés par un utilisateur, ces méthodes de classement sont une partie intégrante de la recherche d'information [87].
Il est possible de répartir les algorithmes de classement en deux grands axes [88]. La classification la plus connue est la classification selon l'importance de la requête d'une liste de documents, ou ils peuvent dépendre de la requête pour classer la liste de documents en fonction de la pertinence entre ces documents et la requête. Nous trouvons aussi, la classification ou une liste de documents est classée selon son importance, sur la base de la technique d'analyse de lien. Par exemple, les modèles de classement actuels de base sont illustrés au figure 5.1.

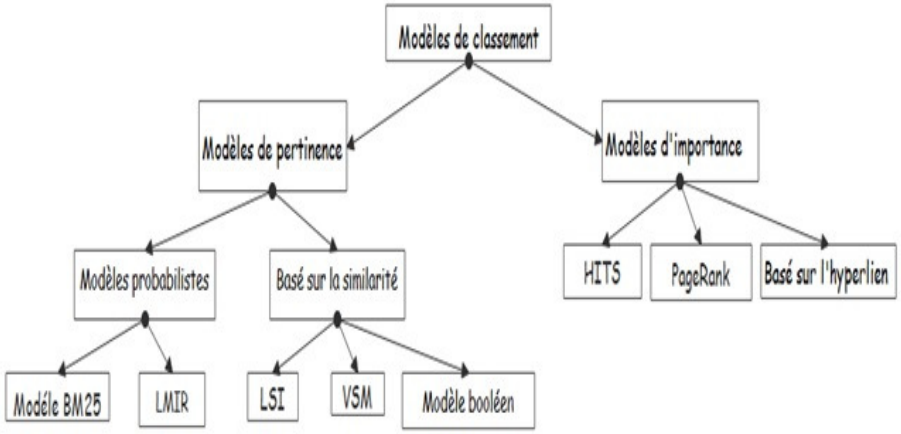

FIGURE 5.1 – *Modèles de classement.*

Comme le montre la figure 5.1, nous avons deux axes, le premier axe représente les modèles de classements de pertinence qui correspondent au classement en fonction de la requête [89]. Parmi ces modèles nous trouvons :
- Le modèle d'espace vectoriel [90] : il permet de représenter des documents texte en tant que vecteurs d'identificateurs. Il est utilisé dans la recherche, le filtrage et le classement par pertinence et index.
- L'analyse sémantique latente [91] : une façon d'analyser la relation entre un ensemble de concepts liés au terme et un ensemble de documents. Cette technique est utilisée dans le traitement du langage naturel.
- L'okapi BM25 [92] : il permet de faire correspondre des documents par une fonction de classement dans les moteurs de recherche, en fonction de leur pertinence avec une requête de recherche donnée. Il est totalement basé sur le cadre de récupération probabiliste.
- Le modèle de classement booléen [93] : une méthode permettant de rechercher la requête de l'utilisateur dans l'ensemble existant, en se basant sur la théorie des ensembles classiques et la logique booléenne.

Le deuxième axe représente un modèle de classement important qui classe selon la technique d'analyse de lien [94]. Parmi ces modèles nous trouvons :
- L'hyperlink-induced topic search (HITS) [95] : un algorithme d'analyse avec comme idée de base est qu'une page web remplit deux fonctions soit fournir des informations et / ou suggérer des liens vers des pages d'un sujet.

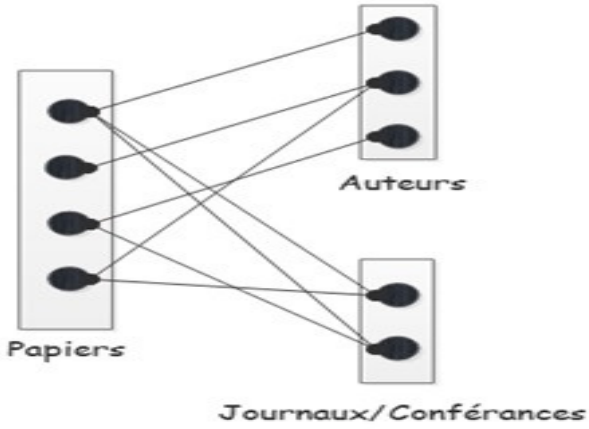

FIGURE 5.2 – *Démonstration de la structure du réseau.*

Comme nous le voyons dans cet exemple proche des publications scientifiques [96] à la figure 5.2, dans le réseau, chaque nœud de papier scientifique est relié à un autre nœud de papier scientifique par des citations entre eux. Ce réseau peut nous aider à tirer plus d'informations sur les auteurs, les articles, les journaux, etc.
- PageRank [97] : un algorithme qui calcule le nombre des liens vers une page ainsi que leurs qualité afin de déterminer l'estimation précise de l'importance des documents pertinents.

La plupart des algorithmes utilisés pour classer les documents de recherche scientifique sont divergents soit du PageRank ou HITS, nous trouvons :
- Topic Rank [98] : regroupe les documents en sujets ; entre les principaux facteurs utilisés, nous trouvons : sujet, citation, date de publication, titre et mot-clé.
- Cite Rank [99] : un algorithme fonctionnant en classant les réseaux de citations en fonction de leur topologie ; entre les principaux facteurs utilisés, nous trouvons : citation, titre et date de publication.
- PTRA [100] : confère un impact plus important à l'âge du papier scientifique et dépend fortement de la date de publication pour classer ces derniers ; parmi les principaux facteurs utilisés, nous trouvons : citation, lieu de publication et date de publication.
- Up Rank, TP Rank [101] : Considère la requête / le sujet, et le contenu comme les axes principaux pour le classement, entre les principaux facteurs utilisés, nous trouvons : citation, mot clé, titre, contenu.
- FutureRank [102] : Estimation du score de prestige futur du classement pour le papier scientifique, entre les principaux facteurs utilisés, nous trouvons : Citation, auteurs et date de publication.

Cependant, tous les chercheurs ont conclu que les deux types d'algorithmes de classement (le pertinent / l'important) présentent certaines limitations, notamment l'algorithme de pertinence qui n'est plus utilisé dans les algorithmes de classement actuel. Cependant,

les études sur la prédiction d'un nouveau rang font encore défaut.
Nous proposons dans la suite, les fondements principaux des méthodes d'apprentissage existant.

5.3 Etat de l'art pour les méthodes d'apprentissage

Dans cette section, nous allons donner une sorte d'aperçu de ce que nous avons trouvé dans la littérature concernant les méthodes d'apprentissage que nous avons utilisées dans nos travaux. Il existe trois familles de méthodes d'apprentissage :
- L'apprentissage supervisé, qui nécessite l'étiquetage préalable des données de classe afin que le modèle puisse s'y former.
- L'apprentissage non supervisé (Clustering), qui ne nécessite pas une saisie préalable d'informations.
- L'apprentissage semi-supervisé, qui manipule conjointement des données non étiquetées et étiquetées.

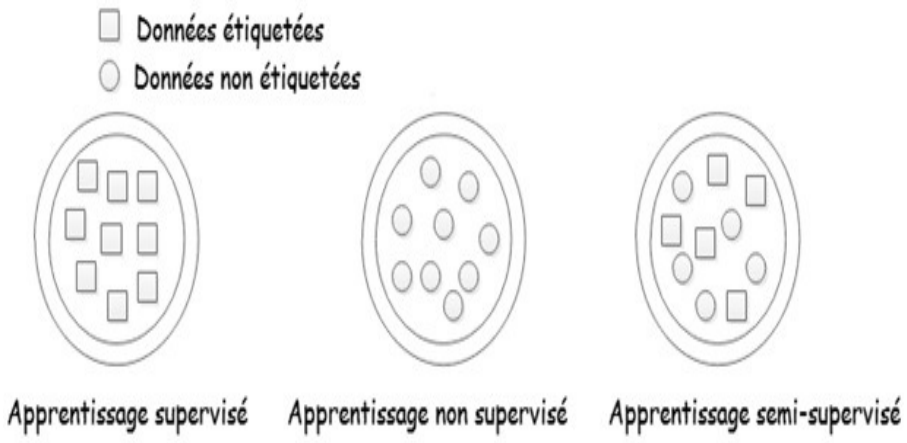

Figure 5.3 – *Sélection des données en fonction de la catégorie d'apprentissage.*

5.3.1 Approche supervisée

L'approche supervisée [103] est une technique d'apprentissage qui conduit automatiquement à la création de règles à partir d'une base de données d'apprentissage contenant des exemples de cas déjà traités. Par conséquent, son objectif est de généraliser, pour les entrées inconnues, qu'il a été en mesure de tirer d'apprentissage des données déjà traitées par des experts ; le but est d'utiliser ceci pour déterminer une représentation compacte de la fonction de prédiction qui, à une nouvelle entrée x, associe une sortie S (x).
Les principaux algorithmes liés à l'apprentissage supervisé sont les suivantes :
- Réseaux de neurones.
- Modèle de markov caché.
- Machines à vecteurs de support.

- Le Boosting.

Les réseaux de neurones [104] sont généralement définis par trois types de paramètres :
- Le modèle d'interconnexion.
- La fonction d'activation.
- Le processus d'apprentissage.

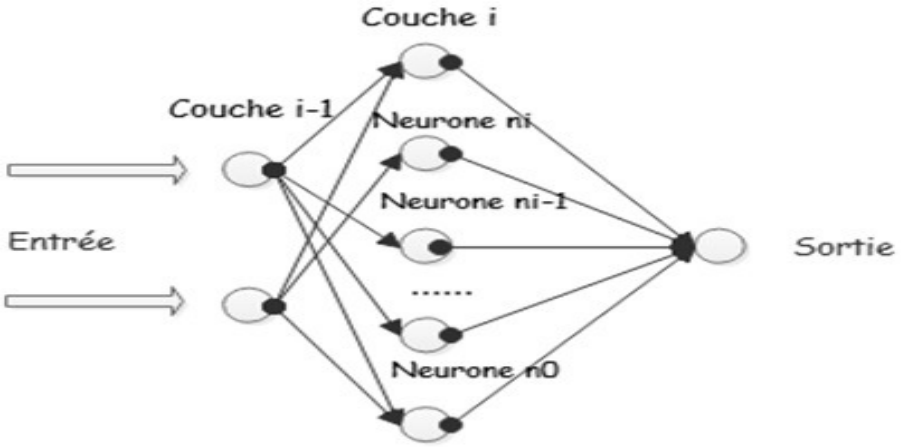

FIGURE 5.4 – *Exemple de réseau de neurones.*

Le modèle de markov caché [105] est défini par deux processus stochastiques : une chaîne de markov est annotée par un ensemble d'états et des transitions entre différents états, appelées probabilités en passant par chaque état. Nous allons mettre en place le processus de décision décrit par :
- Un ensemble fini S d'états discrets notés s.
- Un ensemble fini A d'actions noté a.
- Une fonction de transition P : $S \times A \rightarrow$ P (S), où P (S) est l'ensemble des distributions de probabilité sur S.

La machine à vecteurs de support [106] offre en particulier une bonne approximation du principe fondamental de la minimisation du risque structurel. La méthode dépend des idées suivantes :
- Les données sont projetées dans un grand espace par une transformation basée sur un noyau linéaire, polynôme ou gaussien.
- Les classes sont déconnectées des classifieurs linéaires qui maximisent la marge dans l'espace transformé.
- Les hyper plans peuvent être convertis au moyen de "vecteurs de support".

Le Boosting [107] se résume ainsi :
- Un grand nombre de simples fonctionnalités.

- Pondération d'initialisation pour les ensembles d'entraînement.
- Pour les rondes T :
  — Normalisons les poids.
  — Pour les fonctions de l'ensemble, il faut former un classifieur avec une seule fonction et nous devons examiner l'erreur de formation.
  — Déterminons le classifieur avec l'erreur la plus basse.
  — Mettons à jour les poids des ensembles d'entraînement.
- Le classifieur final est la combinaison linéaire des classifieurs T.

5.3.2 Approche non supervisée

Ce type d'apprentissage peut être motivé par différentes raisons, telles que la charge de développer un étiquetage manuel et la recherche des caractéristiques discriminatoires dans la première étude ou des caractéristiques qui évoluent avec le temps. L'apprentissage non supervisé [108] est souvent traité comme un problème d'estimation de la densité. Les deux principales approches utilisées dans l'apprentissage non supervisé sont les suivantes :
- Le clustering K-Means.
- C-Moyens flous.

Les étapes à suivre pour le clustering k-means [109] se résume ainsi :
- Placez les points K dans l'espace exprimé par les objets groupés.
- Affectez chaque objet au centroïde proche.
- Recalculer les emplacements des K centroïdes.
- Répétez jusqu'à ce que les centroïdes soient réparés, la métrique est ensuite calculée.

C-Moyens flous [110] est très identique à l'algorithme de moyenne, il est résumé comme suit :
- Nommer un certain nombre de groupes.
- Chaque point reçoit un coefficient aléatoire.
- Répétez l'algorithme jusqu'à la convergence.

5.3.3 Approche semi-supervisée

Ce type d'apprentissage consiste à effectuer des tâches génériques d'apprentissage supervisé tout en exploitant des données étiquetées simultanément avec plusieurs données brutes [111]. La première idée consiste à utiliser un contexte non supervisé des sorties prédites par le système lui-même afin de construire les sorties souhaitées en appliquant une technique supervisée, cette approche est connue sous le nom de décision dirigée [112].
La seconde idée dépend de l'utilisation simultanée de deux classifieurs. Ils agissent alternativement en maître-esclave dans un algorithme d'apprentissage itératif : le résultat

calculé par l'un sera considéré comme le résultat approprié par l'autre et réciproquement jusqu'à la convergence [113].
Le critère d'apprentissage est d'optimiser la cohérence entre les deux classifieurs. Cette approche s'appelle l'auto-supervision. Les deux approches majeures de l'apprentissage semi-supervisé sont les suivantes :
- Co-Training [114] : un algorithme d'apprentissage est utilisé lorsqu'il n'y a que des données étiquetées et de grandes quantités de données non étiquetées. L'une de ses utilisations est l'exploration de texte pour les moteurs de recherche.
- Co-Boosting [115] : cela peut être vu comme une combinaison de Co-Training et de Boosting.

Par ailleurs, dans le cadre de cette thèse, nous nous sommes intéressés à la problématique du classement des papiers scientifiques par le biais des algorithmes d'apprentissages supervisés. Le détail de cette contribution sera présenté profondément dans la suite du chapitre.
Nous proposons dans la suite, les fondements principaux de notre méthode de classement des papiers scientifiques basée sur plusieurs paramètres [116].

5.4 Expérience et évaluation

5.4.1 Algorithme de classement proposé

Du point de vue du réseau, les papiers scientifiques peuvent être vus comme des ndans un réseau et les citations entre les papiers scientifiques comme des bords [117].

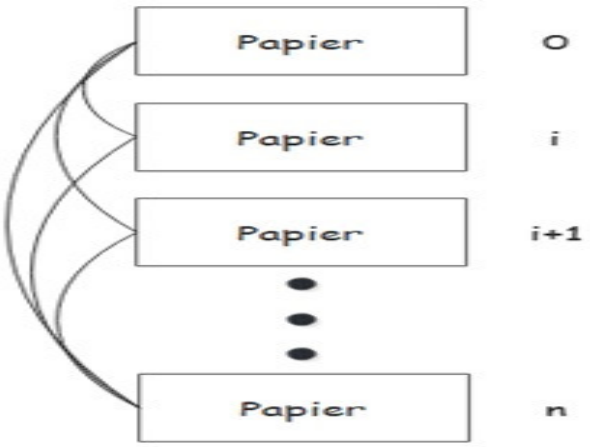

Figure 5.5 – *Le réseau des papiers scientifiques.*

Comme nous voyons à la figure 5.5, dans le réseau, chaque nœud papier X est relié à un autre nœud papier Y par une citation entre eux. Ce réseau peut nous fournit plus d'informations sur les auteurs, les articles, le type d'articles, etc. Ensuite, comme pour tout transfert du modèle, le score d'un papier est calculé en fonction du nombre de citations qui seront transférées aux articles référencés.

De plus, nous devons fractionner nos données en sous-sections pour classer chaque article dans sa division. À titre d'exemple, nous traiterons des informations sur différentes publications de recherche scientifique dans le domaine de l'informatique exclusivement pour le système d'information géographique.
Dans cette directive, nous prenons en compte les points suivant :
- Les papiers scientifiques avec un nombre élevé de citations reflètent l'importance et le prestige de l'auteur.
- Las anciens papiers scientifiques sont toujours au premier classement malgré les dates de publication en raison de ses récentes citations ce qui cause « Scientific Gem » [118].

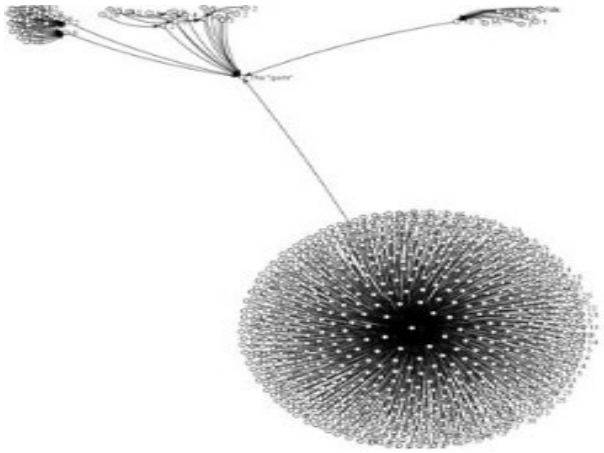

FIGURE 5.6 – *Démonstration de bijou scientifique.*

À partir de la figure 5.6, nous observons clairement que les anciens papiers scientifiques reçoivent un nombre important de citations qui le met en premier classement pour les algorithmes qui se basent sur le nombre de citations, contrairement, des nouveaux papiers scientifiques qui ne reçoivent qu'un nombre limité de citation parmi ces nouveautés apportées dans une discipline donnée.
Les publications récentes ont toujours moins de citations malgré leur plus récente contribution.

Dans tout apprentissage en machine, l'algorithme d'apprentissage détermine les bonnes caractéristiques pouvant donner un meilleur résultat.
Notre méthode de classement est basée sur l'algorithme PageRank avec quelques modifications selon certaines caractéristiques pertinentes [119] ; L'idée clé de notre algorithme de classement dépend de :
- La date de publication du document : nombre d'années écoulées depuis sa publication, selon la formule :

$$A = \text{Année en cours} - \text{Annéede publication}. \tag{5.1}$$

- Le score de la conférence : la qualité de toute conférence peut être explorée en fonction de l'âge de la conférence, de la continuité de la conférence, du nombre de communications dans la procédure et de la bibliothèque numérique impliquée.

- Le score de l'auteur : le nombre d'auteurs de chaque article et le nombre de publications de chaque auteur sont utilisés pour calculer le score de l'auteur, selon la formule :

$$\sum_{i=1}^{n} A_i DL(\frac{1}{NH}) \tag{5.2}$$

- Le taux de téléchargement : le nombre de téléchargements sur le site Web officiel de la revue reflète l'importance que revêt le travail dans le papier de recherche.
- Les mots-clés : l'ordre des mots-clés reflète les sujets et l'intérêt pour le travail.
- Le type de publication : en général, les articles publiés dans les revues scientifiques ont plus d'influence que les autres types de publication ; l'importance de la conférence est moins influente que les revues.
- La publication moyenne / mot clé : Cette fonctionnalité est donnée ci-dessous par :
  - 0 : le mot clé ne figure pas dans le titre du document.
  - 1 : mot-clé figure en résumé.
  - 2 : mot-clé figure dans le titre du papier.
  - 3 : mot clé figure dans le titre et le résumé.
- L'ordre du papier scientifique : peut refléter le degré de perfection d'un travail donné.
- Le nombre de références : dépend de la quantité de littérature disponible sur le sujet.

Le classement du papier scientifique peut être calculé à l'aide de l'équation suivante :

$$Rank = \sum_{i=1}^{n} A_i DL(\frac{1}{NH}) + 0.2(A_p + A_{nbr}) + 0.3(\Delta\omega + Type) + PR(i) \times DL(\frac{1}{1+\log A}) \tag{5.3}$$

Avec :
- Ai : Nombre d'articles publiés par l'auteur.
- N : Nombre total de tous les auteurs pour le papier.
- H : Constante de valeur 10.
- A_p : La publication moyenne / mot-clé.
- A_{nbr} : L'ordre du papier.
- $\Delta\omega$: Taux de téléchargement.
- Type : Le type de papier.
- PR (i) : Le score calculé par l'algorithme PageRank.
- log (A) : Utilisé pour réduire l'impact du vieux papier contenant le plus grand nombre de citations, appelé « Scientific Gem ».

Notre code proposé pour le classement des papiers scientifiques est le suivant :

Procedure: : **Ranking**
Required:
T_i = *Title*
AN = *Authors Number of Works*
N = *Number of Authors for each Paper*
AP = *Date of Publication*
PR = *PageRank Score*

1: For each paper in dataset.
2: Initialize AN, D, PR to 0.0;
3: Get
N[*current paper*],
AN[*current paper*]
AP[*current paper*], PR[*current paper*]
4: compute $AU = (AN)/(NH)$
5: computer $A = 2017 - AP$
6: compute Scientific Research Ranking:

$$Rank = \sum_{i=1}^{n} A_i DL(\frac{1}{NH}) + 0.2(A_v + A_{sh_{ar}}) + 0.3(\Delta\omega + type) \pm PR(i) \times DL(\frac{1}{1 + \log A})$$

end

Pour tester notre algorithme, nous l'appliquons sur les papiers scientifiques relevant de la discipline « système d'information géographique (SIG) ».

5.4.2 Prétraitement des données

Nous attachons une grande importance à la construction de données de formation et de données de test en raison de leur influence sur notre expérience dans ce travail ; pour cette raison, la stratégie de fenêtre mobile [120] a été adoptée.
La taille appropriée des données de test doit respecter certaines conditions. Tout d'abord, il ne doit pas être trop petit car il convergera facilement ; cela aura des conséquences sur l'exactitude de la prévision car, d'une part, il n'a pas fourni suffisamment de données pour appuyer suffisamment les raisons et, d'autre part, il ne doit pas être trop grand, car il n'est pas nécessaire de converger.
Nous avons choisi la stratégie des fenêtres mobiles par rapport à la stratégie des fenêtres coulissantes [121] car la prédiction dépend du facteur temps, ce qui est le cas dans notre étude. Après la réalisation de quelques expériences préliminaires, nous essayons de choisir le modèle de deux données (données de test et données de formation) qui garantissent le taux d'erreur le plus bas possible.
Dans cette recherche, nous avons exploité des données bibliographiques de Thomson-Reuters Web of Science [122], qui contient des informations sur différentes publications de recherche dans différents domaines scientifiques, des métadonnées de base pour les papiers de recherche scientifique dans les domaines de recherche scientifique de 1996 jusqu'en 2015. Il est indispensable de noter que l'ensemble des données Thomson-Reuters que nous avons utilisé contenait des informations sur les citations jusqu'en 2015 uniquement.
Les données tirées du Web of Science contient également des informations qui ne sont pas

utiles dans notre travail. Nous avons besoin de prétraiter l'ensemble des données pour extraire uniquement les informations que nous allons utiliser dans notre traitement.

PT	AU	BA	BE	GP	AF	BF	CA	TI	SO	SE	BS	LA	DT
J	Wilson, MW				Wilson, Matthew W.			On the critic	LANDSCAPE AND URBAN PLANNING			English	Article
J	Mukherjee, F				Mukherjee, Falguni			Public Partic	Geography Compass			English	Article
J	Asare-Kyei, D; Forkuor, G; Venus, V				Asare-Kyei, Daniel; Forkuor, Gerald; Ve			Modeling Fl	WATER			English	Article
J	Brown, G; Fagerholm, N				Brown, Greg; Fagerholm, Nora			Empirical PP	ECOSYSTEM SERVICES			English	Review
J	Lombard, A				Lombard, Andrea			Using partici	JOURNAL OF ENERGY IN SOUTHERN AFI			English	Article
J	Pozzebon, M; Rozas, ST; Delgado, NA				Pozzebon, Marlei; Rozas, Sonia Tello; D			USE AND CO	RAE-REVISTA DE ADMINISTRACAO DE E			English	Article
J	Levine, AS; Feinholz, CL				Levine, Arielle Sarah; Feinholz, Christi			Participatory	APPLIED GEOGRAPHY			English	Article
J	Chingombe, W; Pedzisai, E; Manatsa, D; Mukwada, C				Chingombe, W.; Pedzisai, E.; Manatsa,			A participatc	ARABIAN JOURNAL OF GEOSCIENCES			English	Article
J	McCall, MK; Martinez, J; Verplanke, J				McCall, Michael K.; Martinez, Javier; Ve			Shifting Bou	ACME-AN INTERNATIONAL E-JOURNAL			English	Article
J	Thompson, MM				Thompson, Michelle M.			Public partic	INTERNATIONAL JOURNAL OF DATA MI			English	Article
J	Mekonnen, AD; Gorsevski, PV				Mekonnen, Addisu D.; Gorsevski, Pece			A web-basec	RENEWABLE & SUSTAINABLE ENERGY R			English	Review
J	Cavallo, S; Lynch, J; Scull, P				Cavallo, Sara; Lynch, Joann; Scull, Pete			The Digital D	JOURNAL OF URBAN TECHNOLOGY			English	Article
J	Resch, B; Sagl, G; Tornros, T; Bachmaier, A; Eggers, JI				Resch, Bernd; Sagl, Guenther; Toernro			GIS-Based Pl	ISPRS INTERNATIONAL JOURNAL OF GE			English	Review
J	Al-Wadaey, A; Ziadat, F				Al-Wadaey, Ahmed; Ziadat, Feras			A participatc	JOURNAL OF MOUNTAIN SCIENCE			English	Article
J	Brown, G; Donovan, S; Pullar, D; Pocewicz, A; Toohe				Brown, Greg; Donovan, Shannon; Pulla			An empirical	APPLIED GEOGRAPHY			English	Article
J	Dorn, H; Vetter, M; Hofle, B				Dorn, Helen; Vetter, Michael; Hoefle, B			GIS-Based Rc	REMOTE SENSING			English	Article
J	Sui, D				Sui, Daniel			Opportunitie	TRANSACTIONS IN GIS			English	Review
J	Brown, G; Kelly, M; Whitall, D				Brown, Gregory; Kelly, Maggi; Whitall,			Which 'publi	JOURNAL OF ENVIRONMENTAL PLANN			English	Article
S	Zhang, CX; Yue, P; Zhai, X		IEEE		Zhang, Chenxiao; Yue, Peng; Zhai, Xi			Discovering	THIRD INTER	International Conference		English	Proceedings
J	Brown, G; Schebella, MF; Weber, D				Brown, Greg; Schebella, Morgan Faith;			Using partici	LANDSCAPE AND URBAN PLANNING			English	Article

FIGURE 5.7 – *Extrait des données avant le prétraitement.*

Lors du prétraitement des données, après l'extraction et la préparation des données, la version finale contient : nom du papier de recherche scientifique, auteurs, mots-clés, date du publication, revue/conférence de publication, nombre de citations, taux de téléchargement, publication moyenne / mot-clé, ordre du papier scientifique publié par le chercheur.

On the criticality of mapping practices: Geodesign as critical GIS?	Wilson, MW	In recent yea	Geodesign; (	2015	73	0	8360
Public Participatory GIS	Mukherjee,	Public Participatory GIS (F		2015	84	0	3485
Modeling Flood Hazard Zones at the Sub-District Level with the Rational Model Integrated with GIS and Remote	Asare-Kyei,	Robust risk a community;		2015	78	0	7458
Empirical PPGIS/PGIS mapping of ecosystem services: A review and evaluation	Brown, G; Fa	We review p	Best practice	2015	93	4	9649
Using participatory GIS to examine social perception towards proposed wind energy landscapes	Lombard, A	Thirteen ons	wind energy	2015	30	0	2459
USE AND CONSEQUENCES OF PARTICIPATORY GIS IN A MEXICAN MUNICIPALITY: APPLYING A MULTILEVEL FRAME	Pozzebon, M	This paper s	Participatory	2015	39	0	7561
Participatory GIS to inform coral reef ecosystem management: Mapping human coastal and ocean uses in Hawai	Levine, AS; F	Sociospatial	PGIS; Coral r	2015	52	1	4567
A participatory approach in GIS data collection for flood risk management, Muzarabani district, Zimbabwe	Chingombe,	Recent atter	Chadereka; F	2015	39	1	2389
Shifting Boundaries of Volunteered Geographic Information Systems and Modalities: Learning from PGIS	McCall, MK;	This paper develops a fra		2015	101	0	4589
Public participation GIS and neighbourhood recovery: using community mapping for economic development	Thompson, I	In 2005, New geographic i		2015	24	0	7956
A web-based participatory GIS (PGIS) for offshore wind farm suitability within Lake Erie, Ohio	Mekonnen, A	This study pr	PGIS; Spatial	2015	91	1	1456

FIGURE 5.8 – *Extrait des données après le prétraitement.*

5.4.3 La prédiction du classement du papier scientifique

Pour notre expérience, nous avons divisé nos données en deux sections : la première concerne les données antérieures à 2012 qui représentent notre ensemble de données de formation qui nous permets de calculer le classement des papiers scientifiques ; la deuxième section contient les données d'évaluation que nous souhaitons prédire leur futur classement.

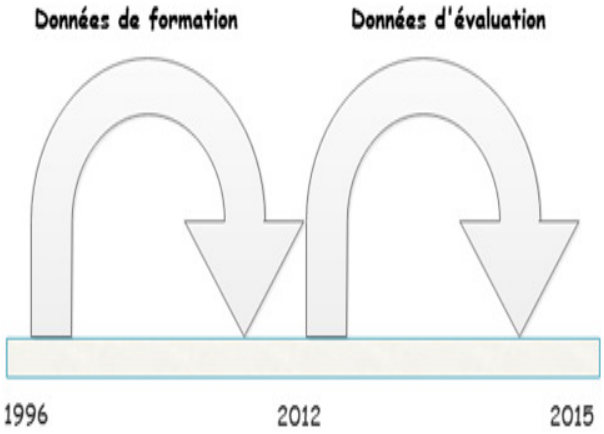

FIGURE 5.9 – *Les données de test ayant débuté avant 2012 et les données d'évaluation ayant été créé après 2012.*

Du point de vue des chercheurs qui cherchaient en 2012 des papiers de recherche scientifique, les seules données disponibles sont les données scientifiques, qui correspondent aux papiers scientifique avant 2012. Notre système peut prévoir pour notre utilisateur les papiers scientifiques qu'il souhaite obtenir avec le meilleur classement dans son domaine de recherche.
Ces expériences ont été conçues pour rechercher les indicateurs pertinents, tels que : identification du papier, score de l'auteur, nombre des papiers scientifiques publiés, taux de téléchargement moyen, nombre de citations. Ces métriques peuvent être calculées à partir de nos données de formation.
Dans notre étude, nous appliquons certaines règles :
- Eliminons les données non intégrales.
- Ne classons pas les papiers scientifiques dont l'auteur ne figure pas dans les données de formation.
- Calculons la moyenne du classement qu'obtient tous les papiers scientifiques pour chaque premier auteur.

À partir de nos données, nous avons choisi de travailler sur les papiers de recherches scientifiques relevant du domaine « système d'information géographique (SIG) » dans notre ensemble de données du Thomson Reuters, comme ensemble de données de formation ; les valeurs de classement de notre algorithme proposé sont basées sur trois points principaux et sont décrites dans le tableau qui suit :

- Auteur et titre où sont présentés les noms des auteurs et les titres des papiers de recherches scientifiques.
- Le DOI est une méthode normalisée d'identification permanente d'un objet électronique publié, une sorte de code permanent et unique de chaque papier scientifique [123].
- Revue et année signifient la date de publication du papier scientifique et le journal /conférence de publication.

Rank	Author and Title	DOI	Journal, Year
1	Mori et al.: Innovative methodology of demand responsive approach for large-scale water supply and sewerage/on-site sanitation projects in developing countries using participatory GIS with high resolution satellite imagery	10.2166/aqua.2011.081	JOURNAL OF WATER SUPPLY RESEARCH AND TECHNOLOGY-AQUA, 2011.
2	Thompson: The city of New Orleans blight fight: using GIS technology to integrate local knowledge	dos.org/10.1080/10511482.2011.634427	HOUSING POLICY DEBATE,2012.
3	Spiegel et al.: Mapping Spaces of Environmental Dispute: GIS, Mining, and Surveillance in the Amazon	10.1080/00045608.2011.641861	ANNALS OF THE ASSOCIATION OF AMERICAN GEOGRAPHERS, 2012.
4	Anderson et al.:Lessons for PPGIS from the application of a decision-support tool in the Nova Forest Alliance of Nova Scotia, Canada	10.1016/j.jenvman.2007.08.031	JOURNAL OF ENVIRONMENTAL MANAGEMENT, 2009.
5	Elwood: Beyond cooptation or resistance: Urban spatial politics, community organizations, and GIS-based spatial narratives	10.1111/j.1467-8306.2006.00480.x	ANNALS OF THE ASSOCIATION OF AMERICAN GEOGRAPHERS, 2006.
6	Krishnamurthy et al.: Mainstreaming local perceptions of hurricane risk into policymaking: A case study of community GIS in Mexico	10.1016/j.gloenvcha.2010.09.007	GLOBAL ENVIRONMENTAL CHANGE-HUMAN AND POLICY DIMENSIONS, 2011
7	Stewart et al.: Public participation geographic information systems (PPGIS): challenges of implementation in Churchill, Manitoba	10.1111/j.1541-0064.2008.00217.x	CANADIAN GEOGRAPHER-GEOGRAPHE, CANADIEN, 2008
8	Brown et al.: Measuring change in place values using public participation GIS (PPGIS)	10.1016/j.apgeog.2011.12.007	APPLIED GEOGRAPHY, 2012
9	Elwood : Geographic Information Science: new geovisualization technologies - emerging questions and linkages with GIScience research	10.1177/0309132508094076	PROGRESS IN HUMAN GEOGRAPHY, 2009
10	Chirowodza et al.: Using Participatory Methods and Geographic Information Systems (GIS) to prepare for an HIV community-based trial in Vulindlela, South Africa (Project Accept-HPTN 043)	10.1002/jcop.20294	JOURNAL OF COMMUNITY PSYCHOLOGY, 2009

FIGURE 5.10 – *Extrait du classement des papiers scientifiques de notre donnée de formation par notre algorithme proposé.*

Maintenant, notre objectif est de prédire le futur classement des papiers de recherche scientifique présentés dans la figure 5.11, qui figurent dans les données d'évaluation dont les auteurs figurent tous dans les données de formation.

```
@relation New-RANK

@attribute Paper-Id
@attribute Title
@attribute Author-score
@attribute Number-paper-published
@attribute Average-Download-rates
@attribute Average-number-of-citation
@attribute Rank{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29}

@data

1,"Participatory GIS to inform coral reef ecosystem management: Mapping human coastal and ocean uses in Hawaii",0.2,10,2389,1,?
2,"Public Participatory GIS",0.1,6,7458,0,?
3,"Modeling Flood Hazard Zones at the Sub-District Level with the Rational Model Integrated with GIS and Remote Sensing Approaches",0.3,7,3649,0,?
4,"Empirical PPGIS/PGIS mapping of ecosystem services: A review and evaluation",0.4,12,2459,4,?
5,"Using participatory GIS to examine social perception towards proposed wind energy landscapes",0.1,9,7561,0,?
6,"USE AND CONSEQUENCES OF PARTICIPATORY GIS IN A MEXICAN MUNICIPALITY: APPLYING A MULTILEVEL FRAMEWORK",0.1,7,4567,0,?
7,"On the criticality of mapping practices: Geodesign as critical GIS?",0.1,5,8560,0,?
8,"A participatory approach in GIS data collection for flood risk management, Muzarabani district, Zimbabwe",0.1,9,7956,1,?
9,"Shifting Boundaries of Volunteered Geographic Information Systems and Modalities: Learning from PGIS",0.099,6,1456,0,?
10,"Public participation GIS and neighbourhood recovery: using community mapping for economic development",0.05,6,4368,0,?
11,"A web-based participatory GIS (PGIS) for offshore wind farm suitability within Lake Erie, Ohio",0.1,10,6319,1,?
12,"The Digital Divide in Citizen-Initiated Government Contacts: A GIS Approach",0.3,5,2996,0,?
13,"GIS-Based Planning and Modeling for Renewable Energy: Challenges and Future Research Avenues",0.8,9,6795,1,?
14,"A participatory GIS approach to identify critical land degradation areas and prioritize soil conservation for mountainous olive groves (case study)",0.1,1,2978,0,?
15,"An empirical evaluation of workshop versus survey PPGIS methods",0.0852,6,6891,0,?
16,"GIS-Based Roughness Derivation for Flood Simulations: A Comparison of Orthophotos, LiDAR and Crowdsourced Geodata",0.1,8,9315,2,?
17,"Opportunities and Impediments for Open GIS",0.1,5,3189,0,?
18,"Which 'public'? Sampling effects in public participation GIS (PPGIS) and volunteered geographic information (VGI) systems for public lands management",0.0175,12,6975,13,?
19,"Discovering Spread Mode of Public Opinions in Incidents and Mapping It with GIS: a Case on Big Geospatial Data Analytics",0.15,7,3915,0,?
20,"Using participatory GIS to measure physical activity and urban park benefits",0.0175,13,9468,10,?
21,"Participatory GIS for strengthening transboundary marine governance in SIDS",0.1,6,3986,2,?
22,"The Global Landscape of GIS in Secondary Education",0.1,7,3954,1,?
23,"GIS and agent-based models for humanitarian assistance",0.2,9,8156,1,?
24,"The Spatial Politics of Affect and Emotion in Participatory GIS",0.2,6,4896,0,?
25,"A place-based approach to conservation management using public participation GIS (PPGIS)",0.025,9,6891,1,?
26,"A Participatory GIS Solution for Watershed Rehabilitation Project Management in the Changjiang and Pearl River Basins",0.1,2,1978,0,?
27,"Using public participation GIS (PPGIS) on the Geoweb to monitor tourism development preferences",0.025,8,6138,5,?
28,"PARTICIPATORY GIS FOR WATER PROVISION AND COMMUNITY PLANNING - CASE STUDY KOFFIEDRAAL, SOUTH AFRICA",0.2,2,1625,0,?
29,"PARTICIPATORY GIS: EXPERIMENTATIONS FOR A 3D SOCIAL VIRTUAL GLOBE",0.1,1,540,0,?
```

FIGURE 5.11 – *Données pour la prédiction.*

De plus, nous avons comparé différents algorithmes d'apprentissage automatique. Dans notre cas d'étude, nous avons choisi de travailler avec l'approche d'apprentissage supervisé, en particulier avec trois classifieurs :
- Les réseaux de neurones représentés par le classifieur Multicouche de Perceptron [124], les raisons de ce choix sont :

— Présentation d'un lecteur au réseau.
— Comparaison de la sortie réseau avec la sortie ciblée.
— Calcul de l'erreur à la sortie de chaque neurone appartenant au réseau.
— Définition de l'augmentation ou de la diminution nécessaire pour obtenir cette valeur.
— Réglage du poids de chaque connexion à l'erreur locale la plus faible.

- Les raisons du choix du classifieur SMO [125] sont les suivantes :

— Recherche d'un multiplicateur de Lagrange $\alpha 1$ qui ne respecte pas les conditions de Karush - Kuhn - Tucker (KKT) pour le problème d'optimisation.
— Choisir un deuxième multiplicateur $\alpha 2$ et optimiser le couple ($\alpha 1$, $\alpha 2$).
— Répéter les étapes 1 et 2 jusqu'à la convergence.
— Calcul de l'erreur à la sortie de chaque neurone appartenant au réseau.

- Les raisons pour choisir le classifieur Kstar [126] sont les suivantes :

— Kstar fonctionne à la volée, ce qui signifie qu'il n'est pas nécessaire que le graphique soit explicitement disponible et stocké dans la mémoire principale, des portions du graphique seront générées au besoin.
— Kstar peut être guidé à l'aide de fonctions heuristiques.

5.4.4 Evaluation d'une méthode de prédiction

Nous allons étudier dans cette partie comment évaluer la performance d'une méthode de prédiction pour s'assurer de sa capacité à satisfaire nos besoins. Le choix d'une mesure, doit être dépendant du type des données à traiter, et nos intérêts souhaités. En partant du principe que la prédiction est une filiale de recherche d'information [127], l'évaluation de la performance d'une méthode de prédiction peut être manifestée à travers plusieurs mesures. Les paramètres d'analyse utilisés dans cette étude sont présentés comme suit :
- Précision : correspond à la fraction d'instances extraites pertinentes, appelée valeur prédictive positive.
- Rappel : correspond à la fraction des instances récupérées qui sont pertinentes, appelée sensibilité.
- F-mesure : est la moyenne harmonique de précision et de rappel ; en d'autres termes, c'est la mesure qui associe précision et rappel, définie comme suit :

$$F - mesure = 2.\frac{(\text{Prćision}.Rappel)}{(\text{Prćision} + Rappel)} \tag{5.4}$$

- Précision moyenne [128] : correspond à la moyenne de la valeur de précision obtenue pour l'ensemble des k documents après l'extraction du document. Avec Q est $d1 \cdots dm_j$ et R_{jk} est l'ensemble des résultats d'extraction classés du premier résultat jusqu'au document d_k, définie comme suit :

$$MAP_d = \frac{1}{d}\sum_{j=1}^{d}(\frac{1}{m_j})\sum_{k=1}^{m_j}P(R_{jk}) \tag{5.5}$$

- Précision moyenne géométrique moyenne [129] : correspond à la moyenne géométrique des valeurs de précision moyennes pour un système de récupération d'informations sur un ensemble de n sujets de requête ; il est défini comme :

$$GMAP = \sqrt[n]{\prod_{n} AP_n} \tag{5.6}$$

- Gain cumulatif : est la somme des valeurs de pertinence notées de tous les résultats dans une liste de résultats de recherche, où rel_i la pertinence notée du résultat à la position i ; il est défini comme :

$$CG_p = \sum_{i=1}^{p} rel_i \tag{5.7}$$

- Gain cumulatif actualisé [130] :correspond à un rang particulier p ; il est défini comme :

$$DCG_p = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)} \tag{5.8}$$

- Gain cumulatif actualisé normalisé [131] : la longueur des listes des résultats de recherche varie en fonction de la requête ; il est défini comme :

$$NDCG_p = \frac{DCG_p}{IDCG_p} \tag{5.9}$$

- Précision moyenne : résume une courbe de rappel de précision sous forme de moyenne pondérée des précisions obtenues à chaque seuil, où P_n et R_n représentent la précision et le rappel au nième seuil :

$$AP = \sum_{n}(R_n - R_{n-1})P_n \tag{5.10}$$

Après l'application des trois classifieurs Multicouche de Perceptron, SMO, Kstar ; dans les données présentées à la figure 5.11, nous avons obtenu les résultats présentés au tableau suivant :

	Multilayer Perceptron	SMO	Kstar
DCG_{26}	129.302	122.092	125.707
$IDCG_{26}$	162.678	162.677	162.683
$NDCG_{26}$	0.794	0.750	0.7727

TABLE 5.1 – *Performance des trois classifieurs.*

Dans le tableau 5.1, nous pouvons clairement voir que le classifieur Multilayer Perceptron

a le plus grand $NDCG_{26}$ par rapport aux SMO et Kstar, ce qui peut nous garantir de bonnes performances pour la prédiction.

La figure ci-dessous illustre notre réseau de prédiction du futur classement après l'utilisation de classifieur Multilayer Perceptron :

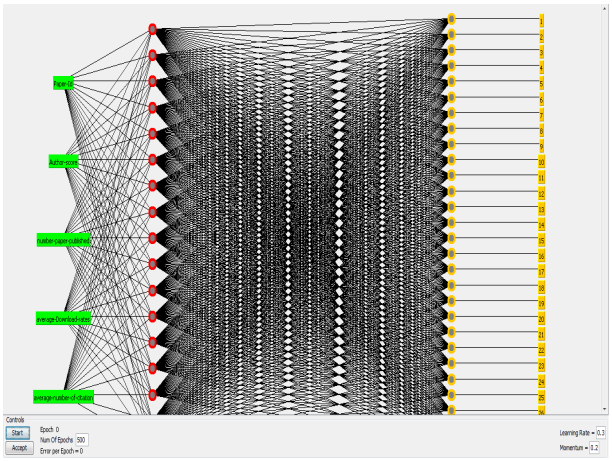

FIGURE 5.12 – *Le réseau de prédiction.*

La figure 5.12 nous présente le réseau de neurones concernant nos données représentées par le classifieur Multilayer Perceptron.
Comme nous voyons à la figure 5.11, notre réseau contient dans les mesures de la couche d'entrée : un identificateur de document, score d'auteur, nombre de publications sur papier, taux de téléchargement moyen et le nombre de citations moyen, et une couche masquée comportant 17 ncomme moyenne entre le nombre d'entrées et de sorties.
Chaque nœud de connexion appelé neurone a un poids calculé à partir de ses entrées avec une fonction sigmoïde [132] :

$$W_{next} = W + \Delta W \tag{5.11}$$

$$\Delta W = -learning\_rate \times gradient + momentum \times \Delta W_{previous} \tag{5.12}$$

Enfin, nous avons la couche de sortie avec 29 classes représentant le futur classement des papiers de recherche scientifique. Les prévisions nous donnent le résultat du futur classement de données d'évaluation qui sont visibles dans le tableau suivant :

Rank	Author and Title	DOI	Journal, Year
1	Wilson: On the criticality of mapping practices: Geodesign as critical GIS?	10.1016/j.landurbplan.2013.12.017	LANDSCAPE AND URBAN PLANNING, 2015
2	Brown et al.: An empirical evaluation of workshop versus survey PPGIS methods	10.1016/j.apgeog.2014.01.008	APPLIED GEOGRAPHY, 2014
3	Mukherjee: Public Participatory GIS	10.1111/gec3.12223	Geography Compass, 2015
4	Brown and Weber: A place-based approach to conservation management using public participation GIS (PPGIS)	10.1080/09640568.2012.685628	JOURNAL OF ENVIRONMENTAL PLANNING AND MANAGEMENT, 2013
5	Brown and Weber: Using public participation GIS (PPGIS) on the Geoweb to monitor tourism development preferences	10.1080/09669582.2012.693501	JOURNAL OF SUSTAINABLE TOURISM, 2013
6	Al-Wadaey and Ziadat: A participatory GIS approach to identify critical land degradation areas and prioritize soil conservation for mountainous olive groves (case study)	10.1007/s11629-013-2827-x	JOURNAL OF MOUNTAIN SCIENCE, 2014
7	Mekonnen and Gorsevski: A web-based participatory GIS (PGIS) for offshore wind farm suitability within Lake Erie, Ohio	10.1016/j.rser.2014.08.030	RENEWABLE & SUSTAINABLE ENERGY REVIEWS, 2015
8	Young and Gilmore: The Spatial Politics of Affect and Emotion in Participatory GIS	10.1080/00045608.2012.707596	ANNALS OF THE ASSOCIATION OF AMERICAN GEOGRAPHERS, 2013

FIGURE 5.13 – *Extrait du futur classement des papiers scientifiques.*

Maintenant, nous analysons les paramètres utilisés pour prédire notre futur classement. Nous proposons quatre variantes APC, APD, ADC et PDC expliquées comme suite :
- Nouveau classement (APC) : variante proposée dans laquelle les paramètres utilisés sont : id-papier, score de l'auteur, nombre de publications et moyenne-citation.
- Nouveau classement (APD) : variante proposée basée sur les mêmes paramètres que APC, mais à la place de la moyenne-citation, nous utilisons le taux de téléchargement moyens.
- Nouveau classement (ADC) : variante proposée dans laquelle nous utilisons : id-papier, le score de l'auteur, le taux de téléchargement moyens et la moyenne-citation.
- Nouveau classement (PDC) : variante proposée basée sur les références papier-id nombre de publications, le taux de téléchargement moyens et la moyenne-citation.

Dans la figure ci-dessous, nous résumons les résultats obtenus pour les quatre variantes de notre futur classement proposé.
Afin de définir la variante appropriée de notre futur classement, les trois paramètres : précision, rappel et f-mesure doivent avoir les meilleures valeurs possibles. Comme nous remarquons sur la figure 5.14, la variante APD présente les meilleures valeurs pour les différents paramètres.

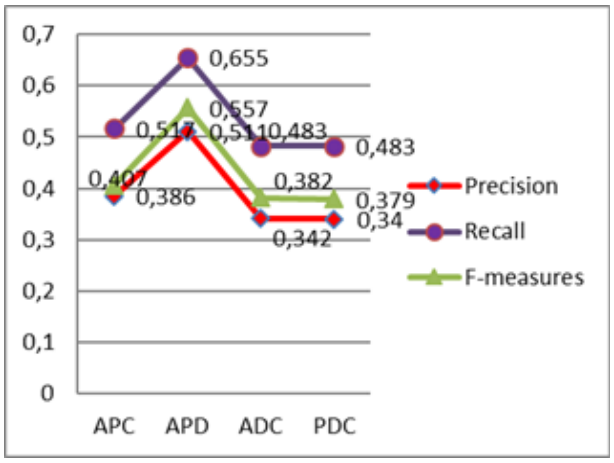

FIGURE 5.14 – *Comparaison de la performance des quatre variantes.*

Comme nous le voyons dans la sous-section précédente, le MAP est une précision moyenne qui est le plus souvent considéré comme une moyenne arithmétique. Contrairement GMAP qu'est une précision géométrique, ce dernier mettait en évidence l'amélioration des sujets peu performants[133].
D'autre part, nous devons comparer MAP et GMAP pour les quatre variantes. Nous voyons clairement sur la figure 5.15. Pour les variantes APD, ADC et PDC les valeurs sont légèrement proches les unes des autres par rapport au variante APC qui a des valeurs supérieures pour MAP et GMAP.

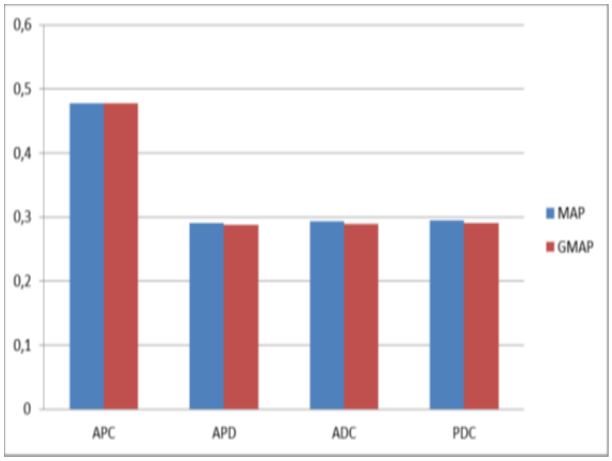

FIGURE 5.15 – *MAP vs GMAP dans les quatre variantes.*

Dans la figure 5.16, nous présentons une comparaison entre les valeurs du GMAP et les valeurs du MAP en fonction du futur classement proposé pour notre variante APC. Nous pouvons constater, pour chaque papier scientifique classé, que le GMAP et le MAP sont proches, sauf pour quelques cas dans lesquels nous trouvons que les valeurs du GMAP sont très inférieures à celles du MAP.

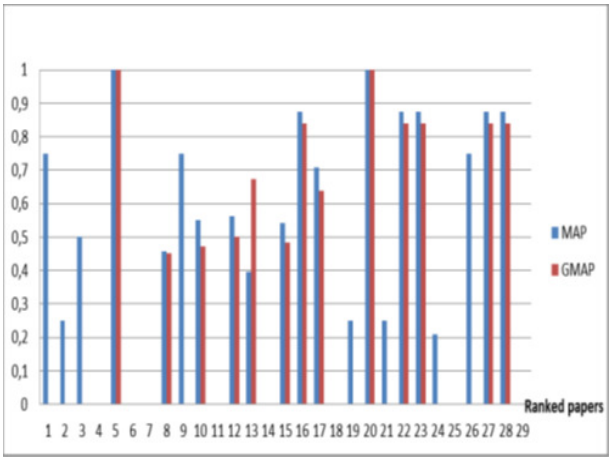

FIGURE 5.16 – *Valeurs des GMAP vs MAP dans le nouveau classement.*

Le MAP et GMAP peuvent être vus comme des mesures similaires de l'efficacité du classement moyen d'une méthode. En résumé, les figures 5.15 et 5.16 montrent que les valeurs du GMAP sont inférieures au MAP, ce qui conduit à un classement performant tout en réduisant les erreurs.
Nous proposons une nouvelle approche pour prédire le futur classement des papiers de recherche scientifique. Notre évaluation expérimentale a montré l'efficacité de l'utilisation de l'algorithme d'apprentissage supervisé dans la discipline du classement. Nous fournissons un algorithme qui utilise différentes métriques telles que : les scores d'auteur, les nombres des papiers publiés, les moyens de taux de téléchargement, le moyen des citations dans un réseau unique. De plus, nous réalisons une comparaison des métriques à leur capacité de prédiction en choisissant la variante la plus performante.

5.5 CONCLUSION

Dans ce chapitre, nous avons décrit les différents types de méthode de classement des papiers scientifiques avec ces différents types dérivés. Ensuite, nous avons abordé l'état de l'art des méthodes d'apprentissage. Ainsi, nous avons présenté notre algorithme de classement avec ces différentes spécifications.
Nous avons présenté également du prétraitement réalisé sur les données issues du Thomson Reuters afin de garder les données qui sont nécessaires pour l'utilisation par notre algorithme. Nous avons cité également, la démarche utilisée pour la prédiction du nouveau classement des papiers scientifiques. Finalement, nous avons introduit des mesures d'évaluation de la performance des méthodes de prédiction, dont la précision, le rappel, MAP, GMAP, DCG et NDCG, qui sont les plus populaires pour choisir la variante adéquate.
Dans le chapitre suivant nous allons soulever la problématique de la prise de décision pour l'attribution des subventions pour les structures de recherches en appliquant les

dernières directives visant à instaurer la bonne gouvernance dans la gestion des universités.

Chapitre 6

Vers une utilisation de la prise de décision à critères multiples dans la recherche scientifique

6.1 Introduction

La bonne gouvernance [134] nécessite la modernisation et la rationalisation de la décision d'information de gestion [135], en aidant le système existant dans ses aspects administratifs et gestionnaires. La gestion et l'automatisation de la recherche scientifique dans les universités publiques représentent un grand défi pour les universités [136], en particulier pour les décideurs qui doivent trouver des techniques et des solutions adaptées à leurs objectifs et spécifications.

La problématique élaborée dans ce chapitre se situe dans le cadre général du système d'information. En effet, la prise de décision est une étape finale pour un décideur, car elle permet de finaliser un processus qui nécessite de suivre des étapes bien définies. Nous nous intéressons particulièrement à la description des méthodes de prise de décision à critères multiples qui, plus précisément, constitue un atout approprié pour la prise de décision pour notre étude.

Le système de prise de décision est un processus basé sur les meilleures pratiques liées à la stratégie de l'université [137], qui traite des situations complexes d'évaluation, de hiérarchisation des priorités et de sélection. Parce que toutes les informations ne sont pas utiles et que de nombreuses informations ne peuvent pas garantir ce que les décideurs doivent déterminer comme : le problème dégagé, le but et l'objectif de la décision, les critères de la décision et les conséquences de cette décision.

En somme, les développements réalisés dans le cadre de ce chapitre se composent de deux axes principaux. Nous présentons d'abord, les méthodes de prise de décision à critères multiples. Ensuite, nous terminons par l'application de certaines méthodes sur un ensemble de données relatives à la recherche scientifique.

6.2 Les méthodes de prise de décision à critères multiples

La prise de décision à critères multiples (MCDM) [138] est utilisée pour résoudre des problèmes liés à plusieurs critères. La prise de décision à critères multiples est regroupée en deux sous-groupes principaux, prise de décision multi-attributs (MADM) et prise de décision à objectifs multiples (MODM), comme nous montre la figure suivante :

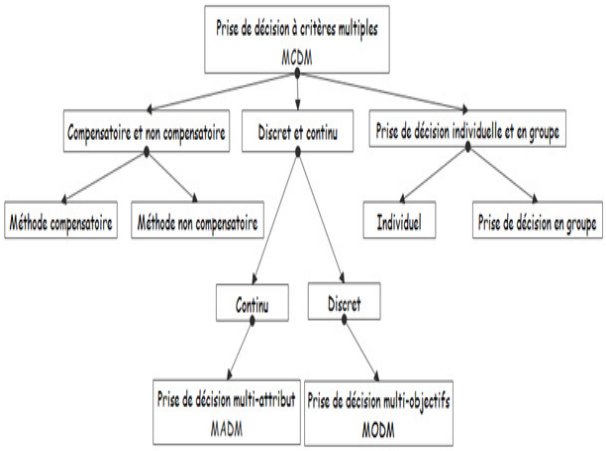

Figure 6.1 – *Catégorie MCDM.*

6.2.1 La prise de décision multi-attributs (MADM)

Premièrement, la prise de décision multi-attributs (MADM) [139] : cette méthode est utilisée pour résoudre des problèmes avec des espaces de décision discrets et un nombre prédéterminé ou limité de choix alternatifs. Cette dernière est lié au jugement de la déclaration personnelle comme au choix de (nouveaux gestionnaires, le choix du nouveau fournisseur...) entre les techniques populaires relevant du MADM, nous trouvons AHP, ANP, TOPSIS, ELECTRE, MAUT et PROMETHEE I & II.

Dans la section suivante, nous décrivons brièvement quelques méthodes de la prise de décision à objectifs multiples.

6.2.2 La prise de décision à objectifs multiples (MODM)

Les méthodes relatives à la prise de décision à objectifs multiples (MODM) : sont utilisée lorsque nous avons des valeurs de variable de décision qui sont déterminées dans un domaine continu ou entier avec un nombre infini ou un grand nombre de choix alternatifs [140], dans le but de satisfaire les contraintes et les priorités de préférence imposées par les décideurs.

6.2.3 Les techniques de prise de décision multicritères

La prise de décision multicritère (MCDM) peut constituer un atout approprié pour la gestion du domaine de la recherche scientifique, les principales étapes suivies dans la prise de décision à critères multiples sont :
- Définissez le problème en spécifiant l'objet qui doit être réaliste et mesurable.
- Déterminez le besoin.
- Établissez les objectifs.
- Identifiez l'alternative.
- Développer des critères d'évaluation.
- Sélection d'outil d'aide à la décision.
- Appliquer l'outil.
- Trouvez le résultat.

Pour la sélection des critères qui doivent être [141] :
- Capable de distinguer les alternatives.
- Assez complet pour couvrir tous les objectifs...
- Non redondant.
- Peu de chiffres.
- Opérationnel et significatif.

6.3 Le principe du processus de hiérarchie analytique (AHP)

Découvert par Saaty le processus de hiérarchie analytique (AHP) [142] a été utilisé pour analyser et structurer des problèmes de décision complexes, le problème de décision est d'abord décomposé en différents critères, la méthode AHP peut être utilisée pour aider les décideurs à calculer le poids de chaque critère en utilisant des jugements de comparaison par paires [143]. Ainsi, elle peut être combiné à une autre technique comme la programmation linéaire, la logique floue... afin de fournir les meilleurs résultats possibles.
L'utilisation de l'AHP est un processus qui comprend les étapes suivantes [144] :
- Définir le problème.
- Déterminer les objectifs et les résultats attendus.
- Déterminez les principaux critères impliqués.
- Hiérarchiser le problème à différents niveaux ; Soit D une matrice de comparaison par paires n x n.

$$D = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix} \tag{6.1}$$

Les éléments diagonaux sont tous égaux à 1.

- Comparez chaque élément du niveau correspondant en normalisant la matrice avec des moyennes géométriques [145] avec :

$$W_i = \frac{[\sum_{j=1}^{n} a_{ij}]^{\frac{1}{n}}}{\sum_{i=1}^{n}[\sum_{j=1}^{n} a_{ij}]^{\frac{1}{n}}} \tag{6.2}$$

- Effectuer un contrôle de cohérence. Si C désigne un vecteur de colonne à n dimensions décrivant la somme de :

$$C = [C_i]_n = DW^T \tag{6.3}$$

avec i = 1,2, ... n

$$DW^T = \begin{bmatrix} 1 & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & 1 \end{bmatrix} = \begin{bmatrix} C_1 \\ C_2 \\ C_n \end{bmatrix} \tag{6.4}$$

- Trouver la valeur propre maximale [146], le taux de consistance (CR) [147] et l'indice de consistance (CI) [148].

$$\lambda_{max} = \frac{\sum_{i=1}^{n} cv_i}{n} \tag{6.5}$$

Où i = 1,2, ... n

$$CI = \frac{\lambda_{max} - n}{n-1} \tag{6.6}$$

$$CR = \frac{CI}{RI} \tag{6.7}$$

Où RI désigne l'indice aléatoire moyen [149].

- Répétez l'opération jusqu'à atteindre les valeurs dans la plage souhaitée.

6.4 Technique pour la préférence de commande par similarité à la solution idéale(TOPSIS)

La technique de préférence par ordre de similarité avec la solution idéale (TOPSIS) a été introduite par Yoon et Hwang en 1981 [150] utilisée dans diverses comparaisons d'alternatives telles que : la sélection des postes de leaders ou d'entités parmi les alternatives, les opérations dans la chaîne logistique [151], l'exploration de données [152], etc.
C'est l'une des méthodes la plus utilisé du MCDM, basée sur la fonction d'agrégation, qui permet de trouver la solution la plus proche de la solution idéale positive et la plus éloignée de la solution idéale négative.
L'utilisation de TOPSIS est organisée à travers les étapes suivant :
- Choisissez une échelle pour mesurer les valeurs du critère.

- Critères alternatifs de matrice X.
- Matrice normalisée par critère (attribut) : nous normalisons tous les scores de la matrice des niveaux attribués aux critères, pour cela, nous appliquons la formule suivante oú x_{ij} critère :

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{n} {x_{ij}}^2}} \tag{6.8}$$

- Matrice normalisée et pondérée : nous multiplions simplement toutes les entrées de la matrice normalisée par la pondération associée à chaque critère :

$$r_{ij} = W_j \times x_{ij} \tag{6.9}$$

- Calcule de la solution favorable idéale A+ : pour chaque critère (attribut), nous calculons la valeur associée la plus favorable A+ en fonction de la nature du critère (favorable ou défavorable) :

$$A^+ = \{max_i x_{ij}(i \in J^+) | min_i x_{ij}(i \in J^-)\} \tag{6.10}$$

- Calcule de la solution défavorable idéale A- : pour chaque critère (attribut), nous calculons la valeur associée la moins favorable A- en fonction de la nature du critère (favorable ou défavorable) :

$$A^- = \{min_i x_{ij}(i \in J^+) | max_i x_{ij}(i \in J^-)\} \tag{6.11}$$

- Calcule de l'écart de la solution favorable idéale de chaque ligne de la matrice :

$$E_i^+ = \sqrt{\sum_{j=1}^{n} ({r_j}^+ - r_{ij})^2} \tag{6.12}$$

- Calcule de l'écart de la solution défavorable idéale de chaque ligne de la matrice :

$$E_i^- = \sqrt{\sum_{j=1}^{n} ({r_j}^- - r_{ij})^2} \tag{6.13}$$

- Calcule du coefficient de proximité [153] de la solution idéale en déterminant son rang dans notre choix qui varie entre 1 et 0, nous choisissons la solution la plus loin possible de la solution la plus idéale défavorable A-, et la plus proche de la solution c'est-à-dire la plus idéale favorable A+ :

$$S_i^* = \frac{E_i^-}{E_i^- + E_i^+} \tag{6.14}$$

Nous proposons dans la suite, l'application des deux techniques déjà abordées dans les sections précédentes, le processus de hiérarchie analytique (AHP) et La technique de préférence par ordre de similarité avec la solution idéale (TOPSIS) sur des données relatives à la recherche scientifique en essayant de répondre au quelques besoins proposés par les décideurs de l'université.

6.5 Résultats et analyse

Nous allons essayer dans cette partie d'appliquer la méthode de prise de décision multi-attributs à l'aide de la technique du processus de hiérarchie analytique (AHP), ainsi qu'avec la technique de préférence par ordre de similarité avec la solution idéale (TOPSIS) sur des données relatives à la recherche scientifique. Dans cette direction, les décideurs veulent connaître le laboratoire qui a obtenu le score le plus élevé l'année dernière en matière de rayonnement scientifique, ainsi que le classement des autres structures de recherche afin d'attribuer des soutiens financiers en fonction des efforts fournis dans la perspective d'appliquer une stratégie de gouvernance sur ce pilier. La décision sera basée sur quatre facteurs :
- Nombre de nouveaux inscrits dans la structure de recherche l'année précédente.
- Nombre de publications scientifiques dans la structure de recherche l'année précédente.
- Nombre de thèses soutenues dans la structure de recherche l'année précédente.
- Nombre d'événements scientifiques organisés par la structure de recherche l'année précédente.

Les décideurs considèrent que le nombre de publications scientifiques des membres de chaque structure de recherche est le facteur le plus important dans la décision, et ils accordent moins d'importance au nombre de nouveaux doctorants inscrits en première année, alors le classement de ce facteur sera :
1) Le nombre des publications scientifiques de l'année précédente.
2) Le nombre de thèses soutenues l'année précédente.
3) Le nombre d'événements scientifiques organisés par le laboratoire l'année précédente.
4) Le nombre de nouveaux inscrits dans le laboratoire l'année précédente.

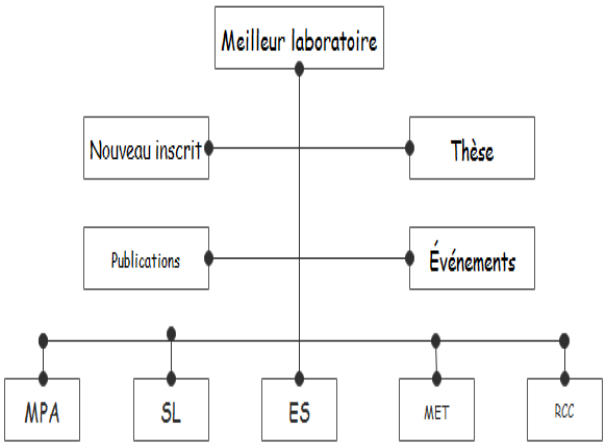

Figure 6.2 – *Schéma des critères de décision et solutions alternatives.*

Avec :
- MPA : Mathématique Physique Appliquée.
- SL : Sciences du Langage.
- ES : Environnement et Santé.

- MET : Modélisation des Ecoulements des Transferts.
- RCC : Recherche Culture et Communication.

Après l'application des étapes exigées par le processus de hiérarchie analytique (AHP), on obtient les résultats suivants :

	Score	Classement
MPA	0.211	3
SL	0.120	5
ES	0.128	4
MET	0.256	2
RCC	0.285	1

TABLE 6.1 – *Classement des structures de recherche en fonction de AHP.*

Sur la base des résultats visionnés du tableau 6.1, nous concluons que la structure de recherche RCC est la structure de recherche scientifique qui correspond le mieux aux critères imposés pour le choix de la structure de recherche idéal suivi du MET, MPA, ES, et SL. L'objectif premier de cette approche est d'aider les décideurs à améliorer la gouvernance universitaire.

Après l'application des étapes exigées par la technique de préférence par ordre de similarité avec la solution idéale (TOPSIS), nous obtenons les résultats suivants :

	Score	Classement
MPA	0.2869	4
SL	0.8536	1
ES	0.2948	3
MET	0.465	2
RCC	0.1686	5

TABLE 6.2 – *Classement des structures de la recherche en fonction du TOPSIS.*

Le classement par ordre décroissant des structures de recherche sur la base des scores et pondérations fournis est le suivant : la structure RCC est la structure qui correspond le mieux au critère imposé pour choisir la structure idéal, suivi de MPA, ES, MET, et SL.

L'évaluation de la performance des deux techniques déjà citées nous fournit les résultats présentés dans le tableau suivant :

	Point minimum	Point maximum	Point moyen
AHP	0.120	0.2895	0.2
TOPSIS	0.1686	0.8536	0.41388

TABLE 6.3 – *Intervalle de point calculé des deux méthodes.*

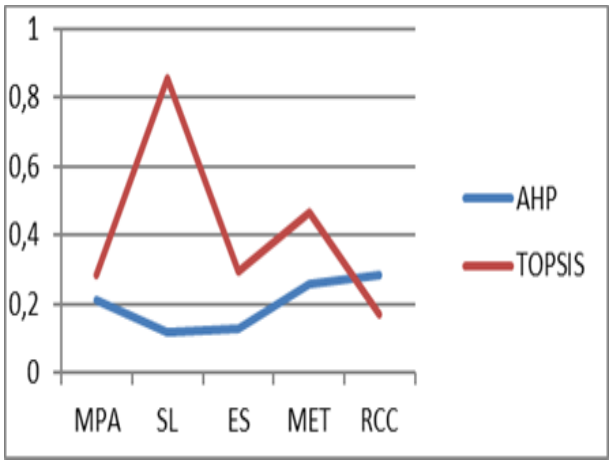

FIGURE 6.3 – *Comparaisons des méthodes d'évaluation.*

Les méthodes AHP et TOPSIS sont appliquées à notre cas avec les mêmes critères, ainsi la distribution de classement des points calculés est indiquée dans le Tableau 6.3. L'intervalle de points TOPSIS est plus élevé que les points AHP. De plus, comme indiqué sur la figure 6.3, la distribution des points calculés avec AHP ne peut pas être distinguée. TOPSIS est meilleur que AHP car les distributions des points calculés avec TOPSIS sont uniformément distinguables plutôt que sur AHP. Par conséquent, la méthode TOPSIS offre les meilleures performances pour l'évaluation, ce qui est déjà déclaré par d'autres études [154].

6.6 CONCLUSION

Dans ce chapitre, nous avons décrit les techniques de La prise de décision à critères multiples (MCDM), ainsi que la prise de décision multi-attributs (MADM), et la prise de décision multi-attributs (MODM). En l'occurrence, nous avons abordé deux techniques de la prise de décision multi-attributs (MADM). En effet, nous avons choisi de travailler avec deux grandes catégories de calcul de similarité : Le processus de la hiérarchie analytique (AHP) et la technique pour la préférence de commande par similarité à la solution idéale (TOPSIS). Ces deux approches présentent néanmoins des caractéristiques complémentaires.

Nous avons présenté également la technique du processus de la hiérarchie analytique (AHP) en présentant les sept étapes essentielles à respecter. Nous avons cité également, la technique pour la préférence de commande par similarité à la solution idéale (TOPSIS) avec ces propres démarches à suivre. Finalement, nous avons évalué la performance des deux techniques appliquée à notre cas d'étude.

Conclusion générale et perspectives

CONCLUSION GÉNÉRALE

Le développement du système de management de la recherche scientifique représente un enjeu majeur pour l'université. A ce sens, il est nécessaire d'adopter une stratégie d'optimisation afin de garantir aux chercheurs tous les outils nécessaires pour automatiser ces taches d'une façon précise et efficace pour pouvoir les motiver et les guider. D'autres part, l'adaptation de ce système selon les besoins et les préférences des décideurs à travers le tableau du bord est indispensable pour avoir une vision globale et au temps réel pour pouvoir viser les besoins et cibler les anomalies au niveau de la gestion et la planification dans le but d'appliquer une stratégie de gouvernance dans le management de ce palier.

Plusieurs questions ont été posées au début de cette thèse, nous en rappelons ici quelques-unes : qu'est-ce que le management de la recherche scientifique ? comment organiser la gestion de la recherche scientifique au sein d'une université ? comment peut-on prédire le domaine du recherche d'un chercheur à partir de son curriculum vitæ ? comment prédire le classement d'un papier de recherche scientifique parmi un corpus donné ? quelle sont les techniques que ne pouvons utiliser pour gouverner la gestion du soutien financier à la structure de recherche scientifique relevant de l'université ? ces questions ont constitué le fil conducteur de l'ensemble de nos recherches.

Les travaux présentés dans cette thèse se situent dans le cadre des problèmes de gestion et management de la recherche scientifique sur ces différents axes. Nous avons opté pour la conception et l'implémentation d'un système basé sur l'utilisation d'un progiciel de gestion intégré afin de pouvoir proposer un système performant et stable, à partir duquel nous pouvons répondre à un ensemble de besoins des chercheurs relevant de l'université dans la formation de thèse. Dans le cadre du même système, nous nous sommes concentrés sur la création d'un tableau de bord interactif pour les décideurs relatifs au pilier de la recherche scientifique, en proposant une solution du tableau de bord qui donne une vision pertinente et en temps réelle sur toutes les structures de recherches et les centres de recherche doctorale, etc.

Le but de notre travail est de faciliter la tâche de traitement automatique des différentes tâches relatives à la recherche scientifique. Dans cette partie, nous avons présenté une nouvelle approche pour le calcul de la similarité sémantique entre les curriculums vitæ des chercheurs dans un corpus de données. Par ailleurs, nous avons proposé une

méthode de prédiction qui vise essentiellement à prédire le domaine de recherche des chercheur en utilisant des algorithmes d'apprentissage supervisé.

En proposant une méthode de classement des papiers de recherche scientifique qui essaye de valoriser l'apparition de nouveaux papiers de recherche scientifique récent dans un domaine de recherche données. Puis, nous avons présenté une étude comparative pour les trois algorithmes de d'apprentissage supervisé (Multilayer Perceptron, SMO et Kstar) pour construire un modèle prédictif basé sur les réseaux de neurones afin de prédire le futur classement des papiers scientifique. Ce modèle contribue à identifier l'impact de cette nouvelle contribution dans une discipline de recherche donnée dans une base de données scientifique.

Ainsi, nous avons présenté une utilisation des méthodes de la prise de décision à critères multiples (MCDM) pour la recherche scientifique afin d'aider les dirigeants à suivre une démarche basée sur les techniques de la prise de décision multi-attributs (MADM), afin d'appliquer une sorte de gouvernance pour l'attribution du soutien financier à la structure de recherche scientifique relevant d'université. Cette dernière contribue à l'instauration d'une sorte de crédibilité au niveau d'attribution du soutien financier visant à augmenter la compétition et concurrence au niveau du rayonnement scientifique entre les différentes structures de recherche existantes.

Certes, les contributions proposées dans cette thèse sont loin d'être parfais. La partie suivante présentera un ensemble de perspectives qui vont permettre l'amélioration de nos contributions en enrichissant leurs fonctionnalités.

PERSPECTIVES

Le travail réalisé dans le cadre de cette thèse nous ouvre plusieurs perspectives de recherche intéressantes que nous comptons développer :

- Dans une première orientation, il s'agit d'appliquer notre approche dans plusieurs domaines d'études similaires, ainsi étendre nos expérimentations à des bases de données de recherche scientifique au niveau du ministère de l'éducation nationale, de la formation professionnelle, de l'enseignement supérieur et de la recherche scientifique et non seulement à notre université.

- Nous comptons aussi comparer les résultats de l'utilisation d'autres classifieurs relevant de l'apprentissage supervisé pour la prédiction du domaine de la recherche des chercheurs en nous basant sur le traitement automatique des curriculums vitæ relative aux chercheurs.

- Dans une deuxième direction, nous comptons appliquer notre méthode de classement des papiers scientifiques sur l'ensemble des papiers scientifique de l'université et du ministère de l'éducation nationale, de la formation professionnelle, de l'enseignement supérieur et de la recherche scientifique, et de comparer les résultats fournis avec d'autre algorithmes d'apprentissage supervisé.

- Dans une troisième direction, nous comptons analyser la réaction des dirigeants sur l'utilisation des méthodes de la prise de décision à critères multiples(MCDM) pour l'attribution de la subvention financière à la structure de recherche pour l'appliquer à d'autres services au sein de l'université.

- Dans une quatrième direction, nous comptons installer une solution datawarhouse qui contient un datamart de recherche scientifique spécifique au besoin de l'université .

Bibliographie

[1] J Oliveira, J M Souza, R Miranda, S odrigues, V Kawamura, R Martino, C Mello, D Krejci, C E Barbosa, L Maia. *GCC : A Knowledge Management Environment for Research Centers and Universities*. In frontiers of WWW Research and Development, APWeb, 2006.

[2] Academic Ranking of World Universities,
http://www.shanghairanking.com

[3] The Times Higher Education World University Rankings,
https://www.timeshighereducation.com

[4] Scopus,
https://www.scopus.com

[5] Science Citation Index (SCI),
http://mjl.clarivate.com

[6] A S Lee, S. H Geoffrey .
Scientific Basis for Rigor in Information Systems Research . MIS Quarterly, 33(2), 2009.

[7] D Romero, F Vernadat.
Enterprise information systems state of the art : Past, present and future trends . Computers in Industry, 79, 2016.

[8] M Swan.
Blockchain for Business : Next-Generation Enterprise Artificial Intelligence Systems . Advances in Computers, 111 :121-162, 2018.

[9] Pascal-francis,
https://pascal-francis.inist.fr

[10] J Mingers, L Leydesdorff. A review of theory and practice in scientometrics . European Journal of Operational Research, 246(1) : 1-19, 2015.

[11] Rafael Ball.
An Introduction to Bibliometrics New Developments and Trends. . European Journal of Operational Research, 246(1) : 1-19, 2015.

[12] Dominika Tkaczyk, PaweÅ Szostek, Mateusz Fedoryszak, Piotr Jan Dendek, Åukasz Bolikowski.
CERMINE : automatic extraction of structured metadata from scientific literature . International Journal on Document Analysis and Recognition, 18(4) : 317-335, Springer, 2015.

[13] R Huchang, T Ming, L Zongmin, B Lev.
Bibliometric analysis for highly cited papers in operations research and management science from 2008 to 2017 based on Essential Science Indicators. . Omega, In press, 2018.

[14] J Kosten.
A classification of the use of research indicators. . Scientometrics, 108(1) : 457-461, Springer, 2016.

[15] K Benmoussa, M Laaziri, S Khoulji, K M Larbi.
SIMARECH 3 : A New Application for the Governance of Scientific Research. . Transactions on Machine Learning and Artificial Intelligence, 5(1), 263-280, 2017.

[16] A Gombault, O Allal-Chérif, A Décamps.
ICT adoption in heritage organizations : Crossing the chasm. . Journal of Business Research, 69(11) : 5135-5140, 2016.

[17] P Bromiley, M McShane, A Nair, E Rustambekov.
Enterprise Risk Management : Review, Critique, and Research Directions . Long Range Planning, 48(4) : 265-276, 2015.

[18] D Romeroa, F Vernadat.
Enterprise information systems state of the art : Past, present and future trends . Computers in Industry, 79 : 3-13, 2016.

[19] A M Aladwani.
Change management strategies for successful ERP implementation . Business Process Management Journal, 7(3) : 66-275, 2006.

[20] F R Jacobs, F C âTed' Weston Jr b.
Enterprise resource planning (ERP)âA brief history.. Journal of Operations Management, 25(2) : 357-363, 2007.

[21] M El Mohadab, B Bouikhalene, S Safi.
Enterprise Resource Planning : Introductory Overview. In 3rd IEEE International Conference on Electrical and Information Technologies, ICEIT, 2017.

[22] P Burcher.
Material Requirements Planning. Wiley Encyclopedia of Management, 2015.

[23] W Howard, P E Oden.
Integrating manufacturing resources planning (MRP II) with flexible manufacturing systems (FMS). Computers Industrial Engineering, 13(1-4) : 107-111, 1987.

[24] A Habadi, Y Samih, E Aljedani.
An Introduction to ERP Systems : Architecture, Implementation and Impacts. International Journal of Computer Applications, 167(9) : 1-4, 2017.

[25] J Ram, M L Wu, R Tagg.
Competitive advantage from ERP projects : Examining the role of key implementation drivers. International Journal of Project Management, 32(4) : 663-675, 2014.

[26] R Malhotra, C Temponi.
Critical decisions for ERP integration : Small business issues. International Journal of Information Management, 30(1) : 28-37, 2010.

[27] B Johansson, F Sudzina.
Choosing Open Source ERP Systems : What Reasons Are There For Doing So ?. IFIP Advances in Information and Communication Technology, vol 299. Springer, Berlin, Heidelberg, 2009.

[28] D L Olson, J Staley.
Case study of open-source enterprise resource planning implementation in a small business. Enterprise Information Systems, 6(1) : 79-94, 2011.

[29] S Nagpal, S K Khatri, A Kumar.
Comparative study of ERP implementation strategies. In Long Island Systems, Applications and Technology, Farmingdale, 2015.

[30] N Pollock, J Cornford.
ERP systems and the university as a âuniqueâ organisation. Information Technology & People, 17(1) : 31-52, 2004.

[31] K Siau, J Messersmith.
Analyzing ERP Implementation at a Public University Using the Innovation Strategy Model. International Journal of HumanâComputer Interaction, 16(1) : 57-80, 2009.

[32] Smile,
`https://www.smile.eu`

[33] J D Rosnay.
Le Macroscope Vers une vision globale. Le seuil, 2014.

[34] M Mohadab, B Bouikhalene, S Safi.
Impact of enterprise resource planning systems On Scientific Research System in Public University. In IEEE The International Arab Conference on Information Technology Yassmine Hammamet, ACIT, 2017.

[35] A Madapusi, D D'Souza.
The influence of ERP system implementation on the operational performance of an organization. International Journal of Information Management, 32(1) : 24-34, 2012.

[36] S K Boell.
Information : Fundamental positions and their implications for information systems research, education and practice. Information and Organization, 27(1) : 1-16, 2017.

[37] B Stvilia, S Wu, D J Lee.
A framework for researcher participation in Research Information Management Systems. The Journal of Academic Librarianship, 45(3) : 195-202, 2019.

[38] M Mitev, K Ilieva,T Apostolov.
Building of scientific information system for sustainable development of BNCT in Bulgaria. Applied Radiation and Isotopes, 67(7-8) : S296-S298, 2009.

[39] Amal Ganesh, K N Shanil, C Sunitha, A M Midhundas.
OpenERP/Odoo - An Open Source Concept to ERP Solution. In IEEE 6th International Conference on Advanced Computing, Bhimavaram, India, IACC, 2016.

[40] H Y Lee, N J Wang.
Cloud-based enterprise resource planning with elastic modelâviewâcontroller architecture for Internet realization. Computer Standards Interfaces, 64 : 11-23, 2019.

[41] J Andress, R Linn.
Introduction to Python. Coding for Penetration Testers (Second Edition), 43-79, 2017.

[42] R E Brun.
XML-based Content Management. Integration, Methodologies and Tools, Chandos Publishing, 2017.

[43] R Daniel.
Odoo development essentials. Chapter 8. Packt Publishing Ltd, 2015.

[44] O Regina ,S H Leo .
PostgreSQL : Up and Running : A Practical Guide to the Advanced Open Source Database. O'Reilly, 2017.

[45] D Mladenić, M Grobelnik.
Automatic Text Analysis by Artificial Intelligence. Informatica, 37 : 27-33, 2013.

[46] J Long.
Useful Process Documents. Process Modeling Style, Chapter 11, 61-65. Morgan Kaufmann, 2014.

[47] K Arlitsch, P Obrien, B Rossmann.
Managing Search Engine Optimization : An Introduction for Library Administrators. Journal of Library Administration, 53(2-3) : 177-188, 2013.

[48] K Ramaboa, P Fish.
Keyword length and matching options as indicators of search intent in sponsored search. Information Processing Management, 54(2) : 175-183, 2018.

[49] N Abid, N Dragan, M L Collard, J I Maletic.
The Evaluation of an Approach for Automatic Generated Documentation. In International Conference on Software Maintenance and Evolution. ICSME, 2017.

[50] C F Reyes, S Shinde.
CV Retrieval System based on job description matching using hybrid word embeddings. Computer Speech Language, 56 :73-79, 2019.

[51] J Tekli, R Chbeir, A J M Traina, C Traina .
SemIndex+ : A semantic indexing scheme for structured, unstructured, and partly structured data. Knowledge-Based Systems, 164 : 378-403, 2019.

[52] R Qu, Y Fang, W Bai, Y Jiang.
Computing semantic similarity based on novel models of semantic representation using Wikipedia. Information Processing Management, 54(6) : 1002-1021, 2018.

[53] N Khasmakhi, M ABalafar, M R Derakhshi.
The state-of-the-art in expert recommendation systems. Engineering Applications of Artificial Intelligence, 82 : 126-147, 2019.

[54] Y Wu, S Xi, Y Yao, F Xu, H Tong, J Lu.
Guiding supervised topic modeling for content based tag recommendation. Neurocomputing, 314(7) : 479-489, 2018.

[55] S Sun, C Luo, J Chen.
A review of natural language processing techniques for opinion mining systems. Information Fusion, 36 : 10-25, 2017.

[56] Y Li, L Yang, B Yang, N Wang, T Wu.
Application of interpretable machine learning models for the intelligent decision. Neurocomputing, 333 : 273-283, 2019.

[57] A Moschitti, R Basili.
Complex Linguistic Features for Text Classification : A Comprehensive Study. In European Conference on Information Retrieval Advances in Information Retrieval, ECIR, 2004.

[58] E Cambria, B White.
Jumping NLP Curves : A Review of Natural Language Processing Research. IEEE Computational Intelligence Magazine, 9(2) : 48-57, 2014.

[59] D A Gachot, E Lange ; J Yang.
The Systran NLP Browser : An Application of Machine Translation Technology in Cross-Language Information Retrieval. Cross-Language Information Retrieval, 2 : 105-118, Springer, 1998.

[60] D Yu, G Hinton, N Morgan, Jen-Tzung Chien, Shigeki Sagayama.
Introduction to the Special Section on Deep Learning for Speech and Language Processing. IEEE Transactions on Audio, Speech, and Language Processing, 20(1) : 4 -6, 2011.

[61] Al Moschitti, R Basili.
Complex Linguistic Features for Text Classification : A Comprehensive Study. In European Conference on Information Retrieval Advances in Information Retrieval, ECIR, 2004.

[62] K v Nunen, J Li, G Reniers, K Ponnet.
Bibliometric analysis of safety culture research. Safety Science, 108 : 248 -258, 2018.

[63] M Mohadab, B Bouikhalene, S Safi.
Automatic CV processing for scientific research using data mining algorithm. Journal of King Saud University - Computer and Information Sciences, 32(5) : 561-567, 2020.

[64] D Belazzouguia, G Navarro, D Valenzuela.
Improved compressed indexes for full-text document retrieval. Journal of Discrete Algorithms, 18 : 3 -13, 2013.

[65] G Chowdhury.
Natural language processing. Annual Review of Information Science and Technology, 37 : 51-89, 2005.

[66] J Giménez, L Màrquez.
Linguistic measures for automatic machine translation evaluation. Machine Translation, 24(3-4), 2010.

[67] K L O'Halloran, S Tan, D S Pham, J Bateman, A V Moere.
A Digital Mixed Methods Research Design : Integrating Multimodal Analysis with Data Mining and Information Visualization for Big Data Analytics. Journal of Mixed Methods Research, 12(1) : 11-30, 2018.

[68] A Ventresque.
Espaces vectoriels sémantiques : enrichissement et interprétationde requÃªtes dans un système d'information distribué et hétérogène.. PhD thesis, Université de Nantes, 2009.

[69] T Sabbah, A Selamat, M H Selamat, F S Al-Anzi, E Herrera, O Krejcar, H Fujita.
Modified frequency-based term weighting schemes for text classification. Applied Soft Computing, 58 : 193-206, 2017.

[70] D Kim, D Seo, S Cho, P Kang.
Multi-co-training for document classification using various document representations : TFâIDF, LDA, and Doc2Vec. Information Sciences, 477 : 15-29, 2019.

[71] J Beel, C Breitinger, S Langer.
Evaluating the CC-IDF citation-weighting scheme â How effectively can 'Inverse Document Frequency' (IDF) be applied to references ?. In the 12th iConference, 2017.

[72] D G Elliman, I T Lancaster.
A review of segmentation and contextual analysis techniques for text recognition. Information Sciences, 23(3-4) : 337-346, 1990.

[73] F N Flores, V Moreira.
Assessing the impact of Stemming Accuracy on Information Retrieval â A multilingual perspective. Information Processing Management, 52(5) : 840 -854, 2016.

[74] L Tao,J Cao, F Liu.
Quantifying textual terms of items for similarity measurement. Information Sciences, 415-416 : 269-282, 2017.

[75] J M L Romera, M M Ballesteros, J G Gutiérrez, J C Riquelme.
External clustering validity index based on chi-squared statistical test. Information Sciences, 487 : 1-17, 2019.

[76] M Ponti, J Kittler, M Riva, T d Campos, C Zor.
On the family of multivariate chi-square copulas. Pattern Recognition, 61 : 470-478, 2017.

[77] R Nisbet, G Miner, K Yale.
The Data Mining and Predictive Analytic Process. Handbook of Statistical Analysis and Data Mining Applications, Second Edition : 39-54, 2018.

[78] N I Karabadji, H Seridi, F Bousetouane, W Dhifli, S Aridhi.
An evolutionary scheme for decision tree construction. Knowledge-Based Systems, 119 : 166-177, 2017.

[79] B Bilalli, A Abelló, T A Banet, R Wrembel.
Intelligent assistance for data pre-processing. Computer Standards & Interfaces, 57 : 101-109, 2018.

[80] A S Altheneyana, M B Menai.
Naïve Bayes classifiers for authorship attribution of Arabic texts. Journal of King Saud University - Computer and Information Sciences, 26(4) : 473-484, 2014.

[81] Z Muda, W Yassin, M N Sulaiman, N I Udzir.
Intrusion detection based on k-means clustering and OneR classification. In 7th International Conference on Information Assurance and Security, IAS, 2011.

[82] V Mitra, C J Wang, S Banerjeec.
Text classification : A least square support vector machine approach. Applied Soft Computing, 7(3) : 908-914, 2007.

[83] S Wu, S McClean.
Performance prediction of data fusion for information retrieval. Information Processing Management, 42(4) : 899-915, 2006.

[84] J Davis, J Davis.
The relationship between Precision-Recall and ROC curves. In 23rd international conference on Machine learning, ICML '06, 2006.

[85] B Twala, M Cartwright.
Ensemble missing data techniques for software effort prediction. Intelligent Data Analysis, 14(3) : 299-331, 2010.

[86] K Sayadi.
Classification du texte numérique et numérisé.Approche fondée sur les algorithmes d'apprentissage automatique. . PhD thesis, Paris, 2017.

[87] T Y Liu.
Learning to Rank for Information Retrieval. Springer-Verlag Berlin Heidelberg, 2011.

[88] T Y Liu.
Learning to Rank for Information Retrieval. Foundations and Trends®in Information Retrieval, 3(3) : 225-331, 2009.

[89] B Long, Y Chang.
Multi-Aspect Relevance Ranking. Relevance Ranking for Vertical Search Engines, 127-145, 2014.

[90] D Yajun, L Wenjun, L Xianjing, P Guoli.
An improved focused crawler based on Semantic Similarity Vector Space Model. Applied Soft Computing,36 : 392-407, 2015.

[91] Z Fuzhen, K George, N Xia, H Qing, S Zhongzhi.
Multi-view learning via probabilistic latent semantic analysis. Information Sciences, 199 : 20-30, 2012.

[92] R Stephen, Z Hugo, T Michael.
Simple BM25 Extension to Multiple Weighted Fields. In International Conference on Information and Knowledge Management, CIKM '04, 2004.

[93] F Lv, H Zhang, J G Lou , S Wang, D Zhang, J Zhao.
CodeHow : Effective Code Search Based on API Understanding and Extended Boolean Model (E). In International Conference on Automated Software Engineering, ASE, 2015.

[94] F Geerts, H Mannila, E Terzi.
Relational link-based ranking. In 30th Annual International Conference on Very Large Data, VLDB, 2004.

[95] L Xinyue, L Hongfei, Z Cong.
An Improved HITS Algorithm Based on Page-query Similarity and Page Popularity. Journal of Computers, 7(1) : 130-134, 2012.

[96] M S Mariania, M Medoa, Y C Zhang.
Identification of milestone papers through time-balanced network centrality. Journal of Informetrics, 10(4) : 1207-1223, 2016.

[97] T Xueyuan.
A new extrapolation method for PageRank computations. Journal of Computational and Applied Mathematics, 313 : 383-392, 2017.

[98] B Adrien, B Florian, D Béatrice.
Topicrank : Graph-based topic ranking for keyphrase extraction. In International Joint Conference on Natural Language Processing, IJCNLP, 2013.

[99] J Pijitra, S Siripun, C Worasit.
CiteRank : combination similarity and static ranking with research paper searching. International Journal of Internet Technology and Secured Transactions, 3(2) : 161-177, 2011.

[100] H M A, L S Feng, H Basheer.
Scientific Research Paper Ranking Algorithm PTRA : A Tradeoff between Time and Citation Network. Applied Mechanics and Materials, 551 : 603-611, 2014.

[101] B S Sohn, J E Jung.
A Novel Ranking Model for a Large-Scale Scientific Publication. Mobile Networks and Applications, 20(4) : 508-520, 2015.

[102] H Sayyadi, L Getoor.
Futurerank : Ranking Scientific Articles by Predicting Their Future PageRank. In International Conference on Data Mining, SIAM, 2009.

[103] D Malte, S H Ulrich, S Balázs.
Supervised learning and Co-training. Algorithmic Learning Theory, 519 : 68-87, 2014.

[104] S Frederic, J Ivan.
An overview of the use of neural networks for data mining tasks. Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery, 2(3) : 193-208, 2012.

[105] Y Zheng, B Jeon, L Sun, J Zhang, H Zhang.
Studentâs t-Hidden Markov Model for Unsupervised Learning Using Localized Feature Selection. IEEE Transactions on Circuits and Systems for Video Technology, 99 : 1-1, 2017.

[106] G S F Javier, Navia Vazquez Ãngel, A M Adrian.
Training Support Vector Machines with privacy-protected data. Pattern Recognition, 72 : 93-107, 2017.

[107] W Y Ng, X Zhou, X Tian, X Wang, D S Yeung.
Baggingâboosting-based semi-supervised multi-hashing with query-adaptive reranking. Neurocomputing, in press, 2017.

[108] J Yao, Q Mao, S Goodison, V Mai, Y Sun.
Feature selection for unsupervised learning through local learning. Pattern Recognition Letters, 53 : 100-107, 2015.

[109] K M Kumar, A R M Reddy.
An efficient k-means clustering filtering algorithm using density based initial cluster centers. Information Sciences, 418-419 : 286-301, 2017.

[110] Z Zarita, P Ong.
An effective fuzzy C-means algorithm based on symmetry similarity approach. Applied Soft Computing, 35 : 433-448, 2015.

[111] X Zhu.
Semi-Supervised Learning. Encyclopedia of Machine Learning, Springer, 892-897, 2011.

[112] C Olivier, B Scholkopf, A Zien.
Semi-supervised learning. IEEE Transactions on Neural Networks, 20(3) : 542-542, 2009.

[113] M R Amini, P Gallinari.
Semi-supervised learning with an imperfect supervisor. Knowledge and Information Systems, 8(4) : 385-413, 2005.

[114] D Malte, S H Ulrich, S Balazs.
Supervised learning and Co-training. Algorithmic Learning Theory, 519 : 68-87, 2014.

[115] S Chen, S Zhu, Y Yan.
Robust visual tracking via online semi-supervised co-boosting. Multimedia Systems, 22(3) : 297-313, 2016.

[116] M Mohadab, B Bouikhalene, S Safi.
Predicting rank for scientific research papers using supervised learning. Applied Computing and Informatics, 15(2) : 182-190, 2018.

[117] G Xu, X Wang, Y Wang, D Lin, X Sun, K Fu.
Edge-Nodes Representation Neural Machine for Link Prediction. Algorithms, 12(1) : 1-16, 2019.

[118] P Chen, H Xie, S Maslov, S Redner.
Finding scientific gems with Googleâs PageRank algorithm. Journal of Informetrics, 1(1) : 8-15, 2007.

[119] M Mohadab, B Bouikhlaene, S Safi.
Towards an Efficient Algorithm for Ranking Scientific Research Papers. In 2nd IEEE international scientific event on internet of things : Recent innovations and challenges, SEIT, 2017.

[120] V Jayaram, L Piyush, P Sachin, B Lorenz.
Development of moving window state and parameter estimators under maximum likelihood and Bayesian frameworks. Journal of Process Control, 60 : 48-67, 2017.

[121] I Chakroun, T Haber, T J Ashby.
The Sliding Window Stochastic Gradient Descent Algorithm. In International Conference on Computational Science, ICCS, 2017.

[122] Thomson-Reuters Web of Science,
`https://clarivate.com`

[123] J Gorraiz, D M Fuentes, C Gumpenberger, J Carlos V Zurian.
The Sliding Window Stochastic Gradient Descent Algorithm. In International Conference on Computational Science, ICCS, 2017.

[124] H Ramchoun, M Amine, J Idrissi, Y Ghanou, M Ettaouil.
Multilayer Perceptron : Architecture Optimization and Training. International Journal of Interactive Multimedia and Artificial Intelligence, 4(1) : 26-30, 2016.

[125] X Shao, K Wu, B Liao.
Single Directional SMO Algorithm for Least Squares Support Vector Machines. Computational Intelligence and Neuroscience, 2013 : 1-7, 2013.

[126] S Lee, M Park, J Park, H Na, M Kwon.
Operator interface programs for KSTAR operation. Fusion Engineering and Design, 88(11) : 2835-2841, 2013.

[127] H Claudia.
Predicting the effectiveness of queries and retrieval systems. In SIGIR Forum, 44(1) : 88, 2010.

[128] M Thelwall.
The precision of the arithmetic mean, geometric mean and percentiles for citation data : An experimental simulation modelling approach. Journal of Informetrics, 10(1) : 110-123, 2016.

[129] S E Robertson.
On GMAP : and other transformations. In International Conference on Information and Knowledge Management, CIKM, 2006.

[130] G Dupret, B Piwowarski.
Model Based Comparison of Discounted Cumulative Gain and Average Precision. Journal of Discrete Algorithms, 18 : 49-62, 2013.

[131] Y Wang, L Wang, Y Li, D He, T Y Liu, W Chen.
A Theoretical Analysis of NDCG Type Ranking Measures. In Conference on Learning Theory, COLT, 2013.

[132] A Iliev, N Kyurkchiev, S Markov.
On the approximation of the step function by some sigmoid functions. Mathematics and Computers in Simulation, 133 : 223-234, 2017.

[133] M Sanderson.
'Test collection based evaluation of information retrieval systems. Foundations and Trends in Information Retrieval, 4(4), 247-375, 2010.

[134] C Wang, R Medaglia, L Zheng.
Towards a typology of adaptive governance in the digital government context : The role of decision-making and accountability. Government Information Quarterly, 38(2) : 306-322, 2018.

[135] G Cao, Y Duan, T Cadden.
The link between information processing capability and competitive advantage mediated through decision-making effectiveness. International Journal of Information Management, 44 : 121-131, 2019.

[136] K Benmoussa, M Laaziri, S Khoulji, M L Kerkeb, A Yamami.
Enhanced model for ergonomic evaluation of information systems : application to scientific research information system. International Journal of Electrical and Computer Engineering, 9(1) : 683-694, 2019.

[137] P Giurin, F Munari, A Scandura, L Toschi.
The strategic orientation of universities in knowledge transfer activities. Technological Forecasting and Social Change, 138 : 261-278, 2019.

[138] A Ishizaka, S Siraj.
Are multi-criteria decision-making tools useful ? An experimental comparative study of three methods. European Journal of Operational Research, 264(2) : 462-471, 2018.

[139] K Y Huang, I Hui Li.
A multi-attribute decision-making model for the robust classification of multiple in-

puts and outputs datasets with uncertainty. Applied Soft Computing, 38 : 176-189, 2016.

[140] A J Nebro, A B Ruiz, C B Gonzalez, J G Nieto, M Luque, J F Aldana Montes.
InDM2 : Interactive Dynamic Multi-Objective Decision Making Using Evolutionary Algorithms. Swarm and Evolutionary Computation, 40 : 184-195, 2018.

[141] R Yager.
Categorization in multi-criteria decision making. Information Sciences, 460â461 : 416-423, 2018.

[142] H William, M Xin.
The state-of-the-art integrations and applications of the analytic hierarchy process. European Journal of Operational Research, 267(2) : 399-414, 2018.

[143] M Hanine, O Boutkhoum, A Tikniouine, T Agouti.
Application of an integrated multi-criteria decision making AHP-TOPSIS methodology for ETL software selection. SpringerPlus, 5(1), 263-280, 2016.

[144] O S Vaidyan, S Kumar.
Analytic hierarchy process : An overview of applications. European Journal of Operational Research, 169(1) : 1-29, 2006.

[145] S Kim, H Lee, Y Lim.
Repetition invariant geometric means. Linear Algebra and its Applications, 544 : 443-459, 2018.

[146] A Chowdhury, L Hogben, J Melancon, R Mikkelson.
Rational realization of maximum eigenvalue multiplicity of symmetric tree sign patterns. Linear Algebra and its Applications, 418(2-3) : 380-393, 2006.

[147] P Chu, J H Liu.
Note on consistency ratio. Mathematical and Computer Modelling, 35(9-10) : 1077-1080, 2002.

[148] G Khatwania, A K Kar.
Improving the Cosine Consistency Index for the analytic hierarchy process for solving multi-criteria decision making problems. Applied Computing and Informatics, 13 (2) : 118-129, 2017.

[149] N Chatterjee, P K Sahoo.
Random Indexing and Modified Random Indexing based approach for extractive text summarization. Computer Speech Language, 29 (1) : 32-44, 2015.

[150] H C Lai, K Y Hwang, C L Yoon.
Multiple attribute decision making : methods and applications a state-of-the-art survey. Springer Science Business Media, 186, 2012.

[151] K Zare, J M Tekmehn, S Karimi.
A SWOT framework for analyzing the electricity supply chain using an integrated

AHP methodology combined with fuzzy-TOPSIS. International Strategic Management Review, 3 (1-2) : 66-80, 2015.

[152] Y Peng, Y Zhang, Y Tang, S Li.
An incident information management framework based on data integration, data mining, and multi-criteria decision making. Decision Support Systems, 51 (2) : 316-327, 2011.

[153] G Dwivedi, R KSrivastava, S K Srivastava.
A generalised fuzzy TOPSIS with improved closeness coefficient. Expert Systems with Applications, 96 : 185-195, 2018.

[154] S H Zanakis, A Solomon, N Wishart, S Dublish.
Multi-attribute decision making : A simulation comparison of select methods. European Journal of Operational Research, 107(3) : 507-529, 1998.

ANNEXE

Chaos, Solitons and Fractals 139 (2020) 110052

Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

Nonlinear Science, and Nonequilibrium and Complex Phenomena

journal homepage: www.elsevier.com/locate/chaos

Bibliometric method for mapping the state of the art of scientific production in Covid-19

Mohamed El Mohadab*, Belaid Bouikhalene, Said Safi

Laboratory of Mathematics Innovation and Information Technology (LIMATI), Department of Mathematics and computers Sciences, Polydisciplinary faculty Beni Mellal, Sultan Moulay Slimane University, Morocco

ARTICLE INFO

Article history:
Received 3 June 2020
Accepted 23 June 2020
Available online 30 June 2020

Keywords:
Covid-19
Scientific production
Bibliometric method
Bibliometric analysis
Scientific research

ABSTRACT

Global scientific production around the Covid-19 pandemic, in the various disciplines on the various international scientific bibliographic databases, has grown exponentially. The latter builds a source of scientific enrichment and an important lever for most researchers around the world, each of its field and its position with an ultimate aim of overcoming this pandemic. In this direction, bibliometric data constitute a fundamental source in the process of evaluation of scientific production in the academic world; bibliometrics provides researchers and institutions with crucial strategic information for the enhancement of their research results with the local and international scientific community, especially in this international pandemic.

© 2020 Elsevier Ltd. All rights reserved.

1. Introduction

The latest statistics indicate that there has been an exponential increase in the number of publications since the discovery of the Covid-19 pandemic; the results provide a comprehensive view of interdisciplinary research in medicine, biology, finance and other fields.

The number of publications in international databases aims to disseminate and share the contributions and advances of academic research from different groups of researchers from different universities and countries in the thematic of Covid-19.

Bibliometrics [1] is a tool for mapping the state of the art in a field related to given scientific knowledge. So the use of bibliometric analysis [2] to identify and analyze the scientific performance of authors, articles, journals, institutions, countries through the analysis of keywords and the number of citations constitutes an essential element which provides researchers with the means to identify avenues and new directions in relation to a theme of scientific research.

2. Bibliometrics at the service of scientific research

Scientometrics [3] is considered as the science of measurement and the analysis of science which is based on an input set and an output set which uses bibliometrics in the field of study of publications. The latter is a meta-science which takes science as its object of study based on three elements of scientific activity: its inputs, its outputs and its impacts. Thus, it makes it possible to map and broaden knowledge on a research field, by clarifying the links between the authors, the publications, the institutions, and other characteristics of the studied field.

Scientific publications [4] represent all publications in newspapers or conferences, either chapters in scientific books or scientific patents. All these types of publications represent the work of a researcher who publishes these works with the aim of circulating these results in databases which have broad international visibility and scientific credibility such as Web of Science, Scopus... and renowned publishing houses such as Elsevier, Springer, Wiley, etc.; but with all the efforts made, the benefits that can be drawn remain limited if we cannot manage this large mass of publication which is added every day to the thousands or millions of existing scientific papers.

Bibliometric data is used for:

- Measure and compare the scientific output of the researcher, research groups, institutions, regions or countries using indicators based on:
  - The number of publications.
  - The quotes received.
  - The collaborations.
- Identify the most important or influential journals in a given field.
- Monitor the evolution over time of a discipline or research subject.

* Corresponding author.
E-mail address: m.elmohadab@gmail.com (M.E. Mohadab).

https://doi.org/10.1016/j.chaos.2020.110052
0960-0779/© 2020 Elsevier Ltd. All rights reserved.

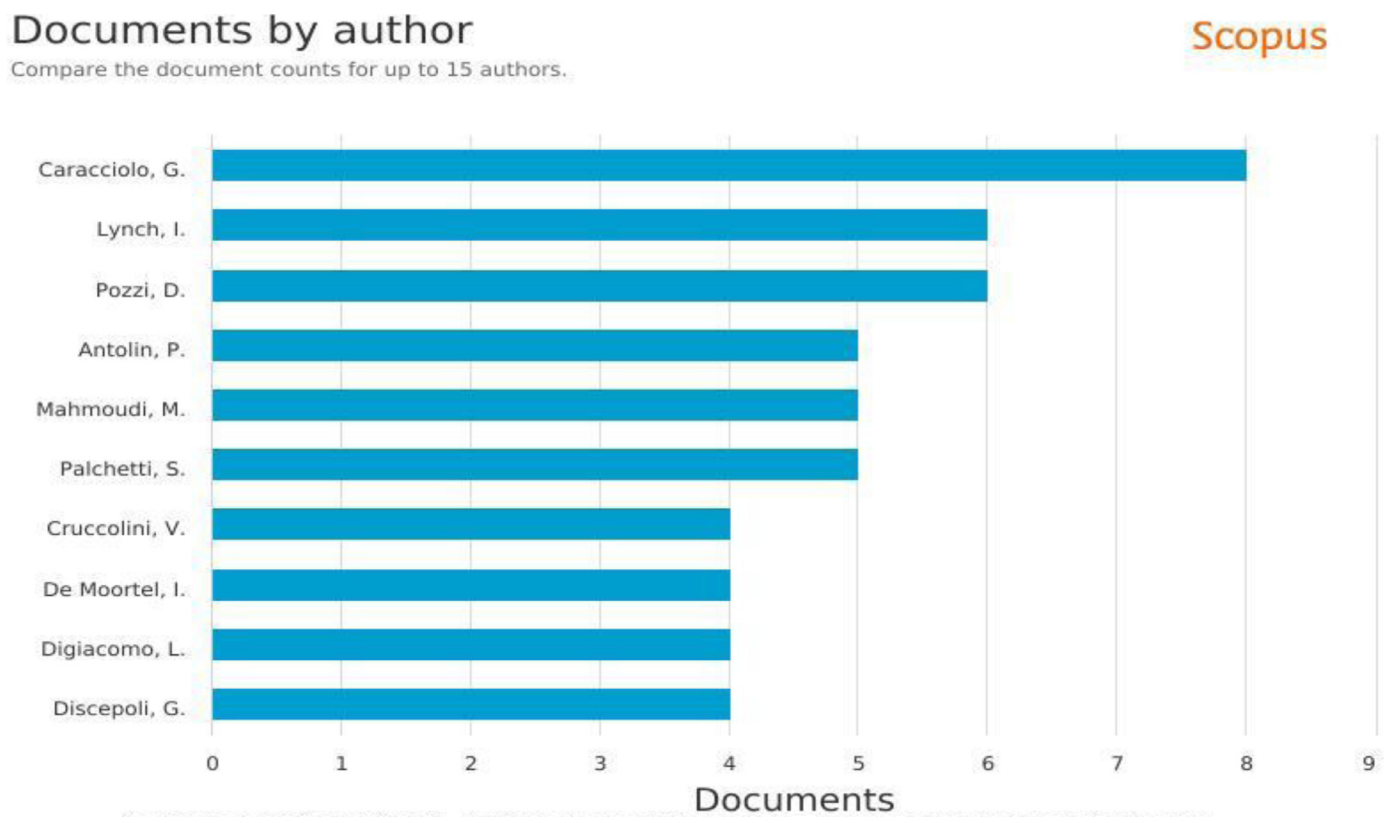

Fig. 1. Statistics of the best author published for Covid-19 on Scopus.

These data represent the main part of the data provided for each paper by the databases which allow bibliometrics to carry out statistical processing, and bibliometric analysis.

3. Statistical overview on Covid-19

3.1. The international context

According to statistics provided by Johns Hopkins University [5] until May 23, 2020, the death of more than 339,949 people worldwide, was the infection of 5,267,452, considerable efforts were made in the various disciplines relating to the treatment of this pandemic either from near or far.

Since the beginning of the year, Covid-19 represents an increasing interest for researchers from all over the world, in response to this crisis, a lot of research was carried out in many fields of research (medical, biology, financial, ...) by several Institutions and organizations, either public or private worldwide, each with their own means available.

By reviewing most of the scientific databases, the search to identify the scientific output related to the subject of Covid-19 [6] was carried out using a set of terms as search criteria, the language of the documents is the English because it is the universal language of research, all disciplines are authorized in order to provide a global view of Covid-19 research in the various disciplines, research is limited to the period from early 2020 (Beginning of the pandemic a been listed) so far Figs. 1–17.

❖ SCOPUS [7]:

Using the Scopus search engine to search for the word “covid-19” and “coronavirus” from 01/01/2020 until 23/05/2020, we find 10,228 documents:

- According to the authors:
- According to the institutions:
- According to the country:
- According to the type of documents:
- According to the domains:

❖ Web of Science [8]:

Using the search engine of Web of Science to search for the word “covid-19” and “coronavirus” from 01/01/2020 until 23/05/2020 results in 5,161 documents:

- According to the authors:
- According to the institutions:
- According to the country:
- According to the type of documents:
- According to the domains:

3.2. The African and Arab context

❖ Scopus:
■ Africa:
■ Arab:
❖ World of Science:
■ Africa:
■ Arab:

4. Methodology of the analysis of bibliometric data

The exploitation of the bibliometric parameters available on the scientific data base on multiple field and discipline makes it possible to release relevant information which can meet the expectations of researchers, research teams and research institutes. The bibliometric analysis reveals to the researcher exact information for the construction of new research as in the case of our study on Covid-19.

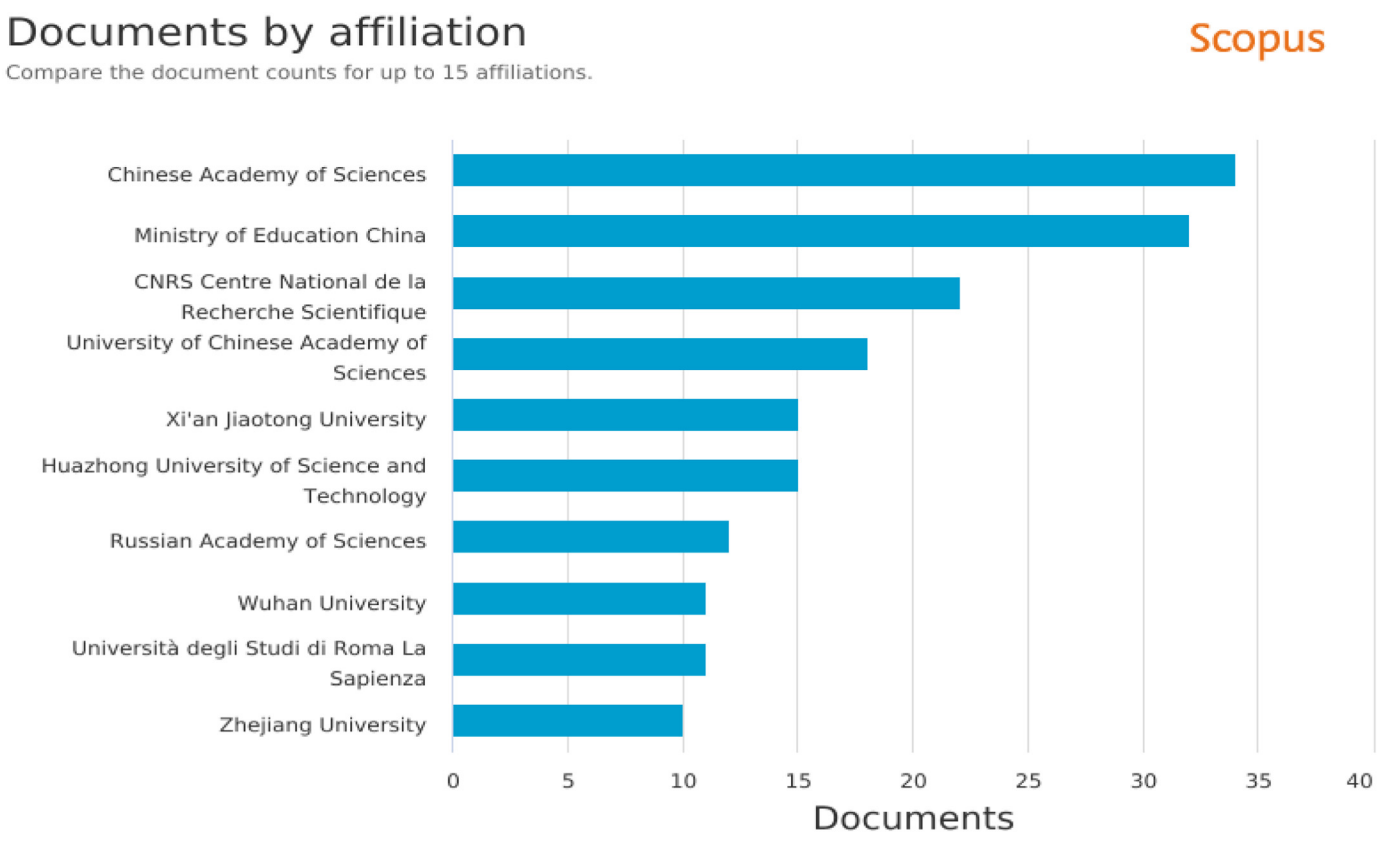

Fig. 2. Statistics of the best 10 institutions published for Covid-19 on Scopus.

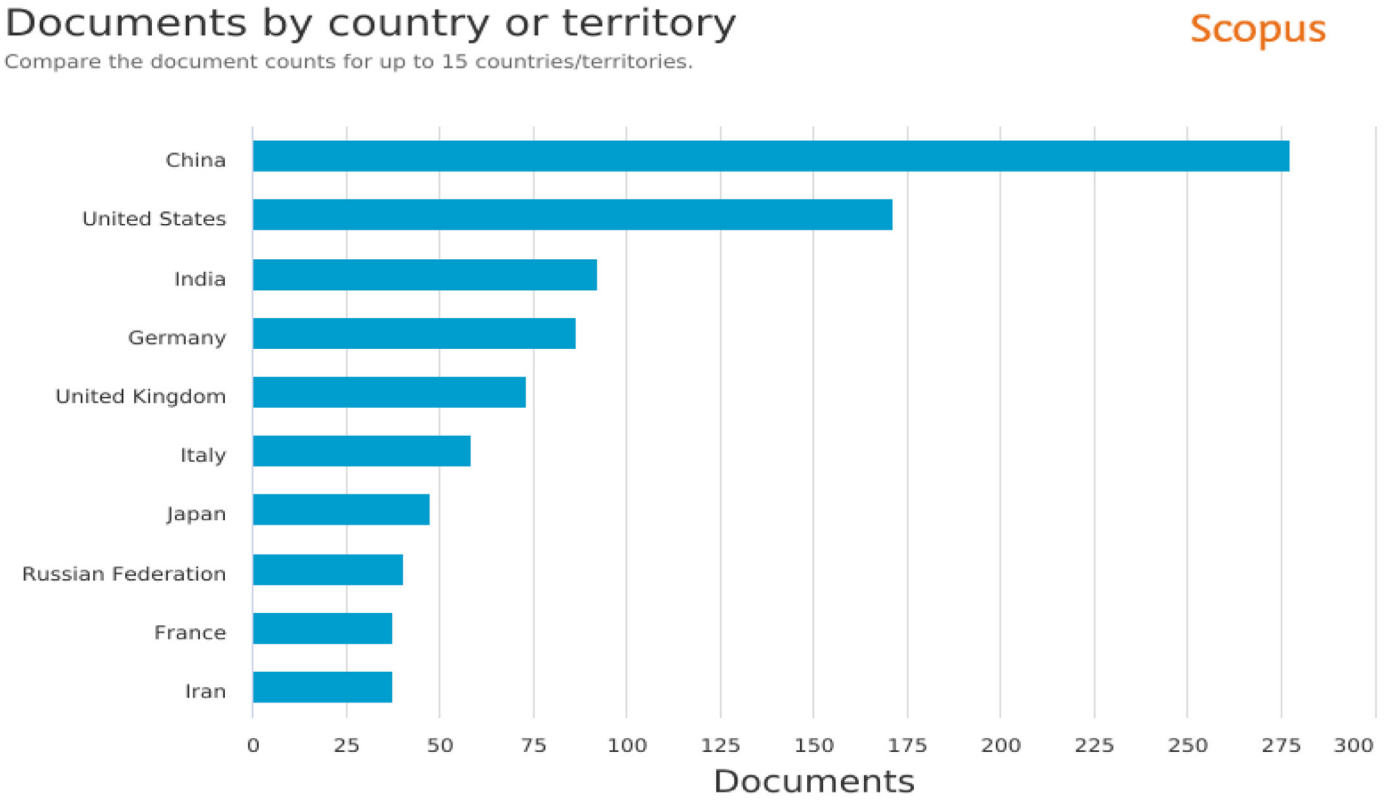

Fig. 3. Statistics of the best 10 countries published for Covid-19 on Scopus.

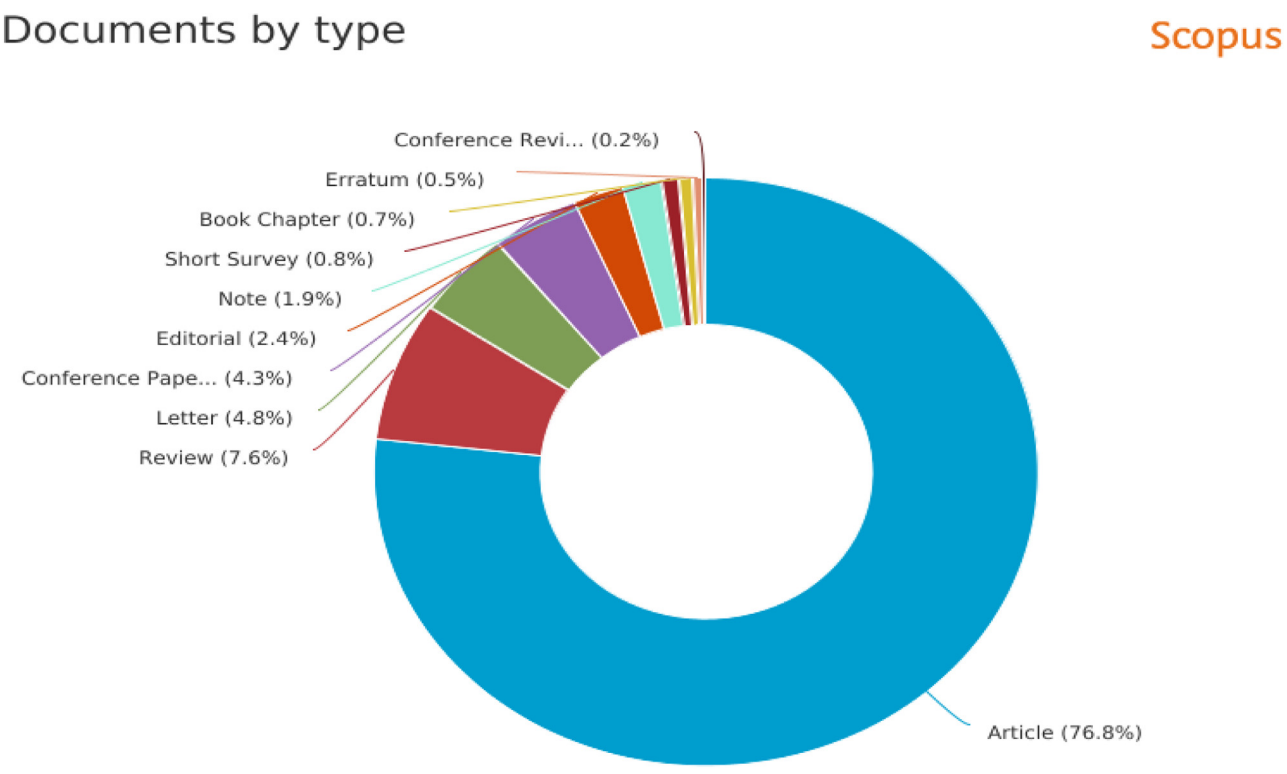

Fig. 4. Statistics of the type of document published for Covid-19 on Scopus.

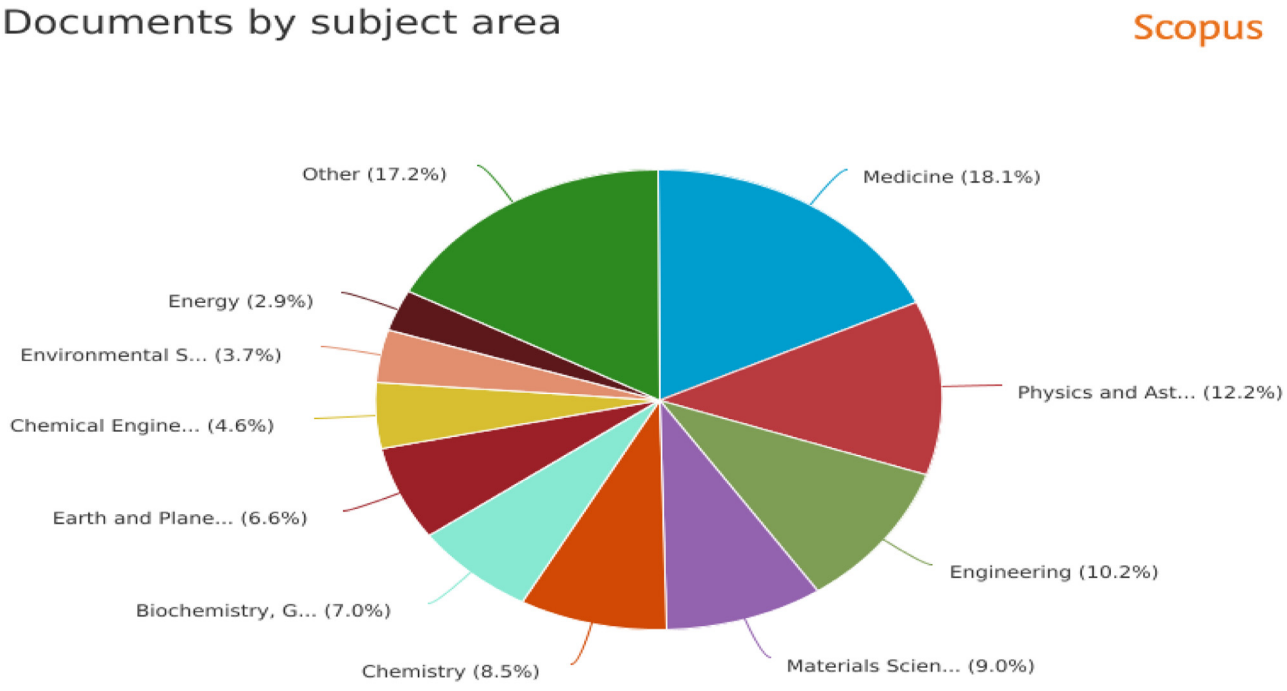

Fig. 5. Domain statistics published for Covid-19 on Scopus.

This study was carried out on the basis of specific research using the three databases (Scopus, Web of Science, Pubmed) from the beginning of 2020 until 23/05/2020. The sample consists of 5,161 academic publications (Web of Science), 10,228 academic publications (Scopus) and 7,991 academic publications (Pubmed). The use of bibliometrics will contribute to the exploration and description of the existing scientific literature on the theme of Covid-19.

The steps taken to achieve the desired results are manifested as:

The use of bibliometric tools plays an important role in guiding a particular field of study by collecting scientific data and synthesizing the results obtained.

Statistics from different bibliographic databases which differ either in terms of data volume or coverage constitutes a reliable source for bibliometric indicators [9].

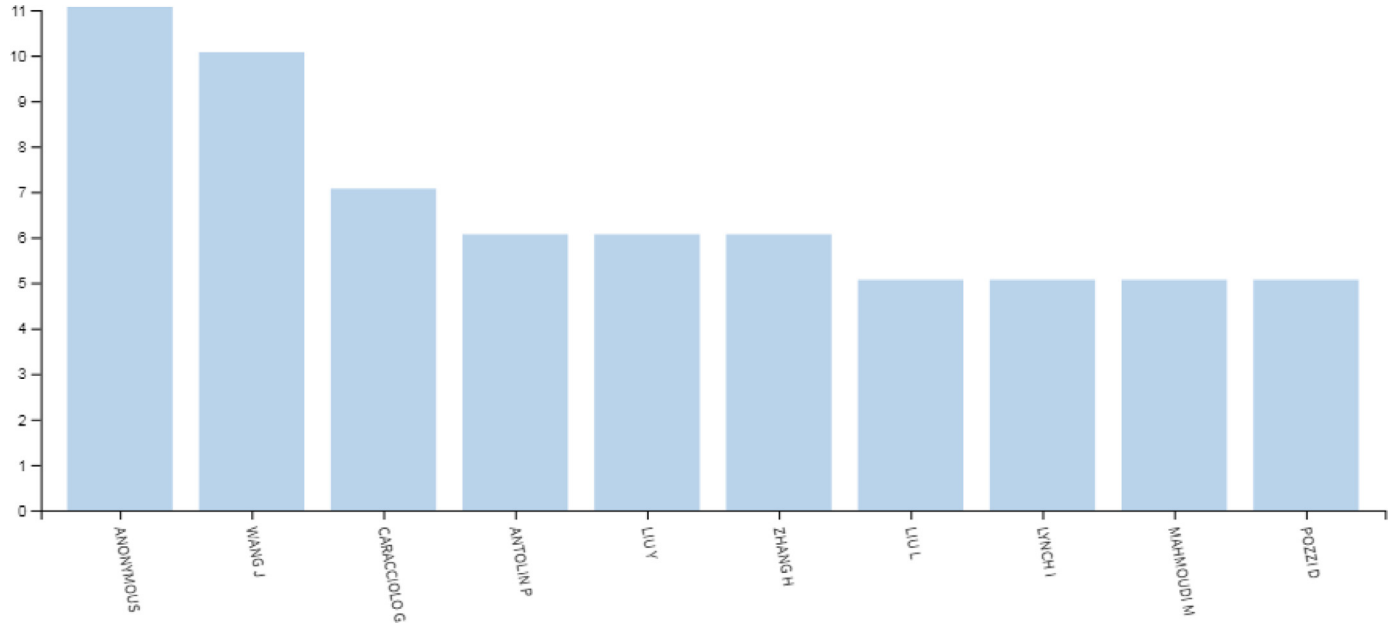

Fig. 6. Statistics of the best 10 author published for Covid-19 on Web of Science.

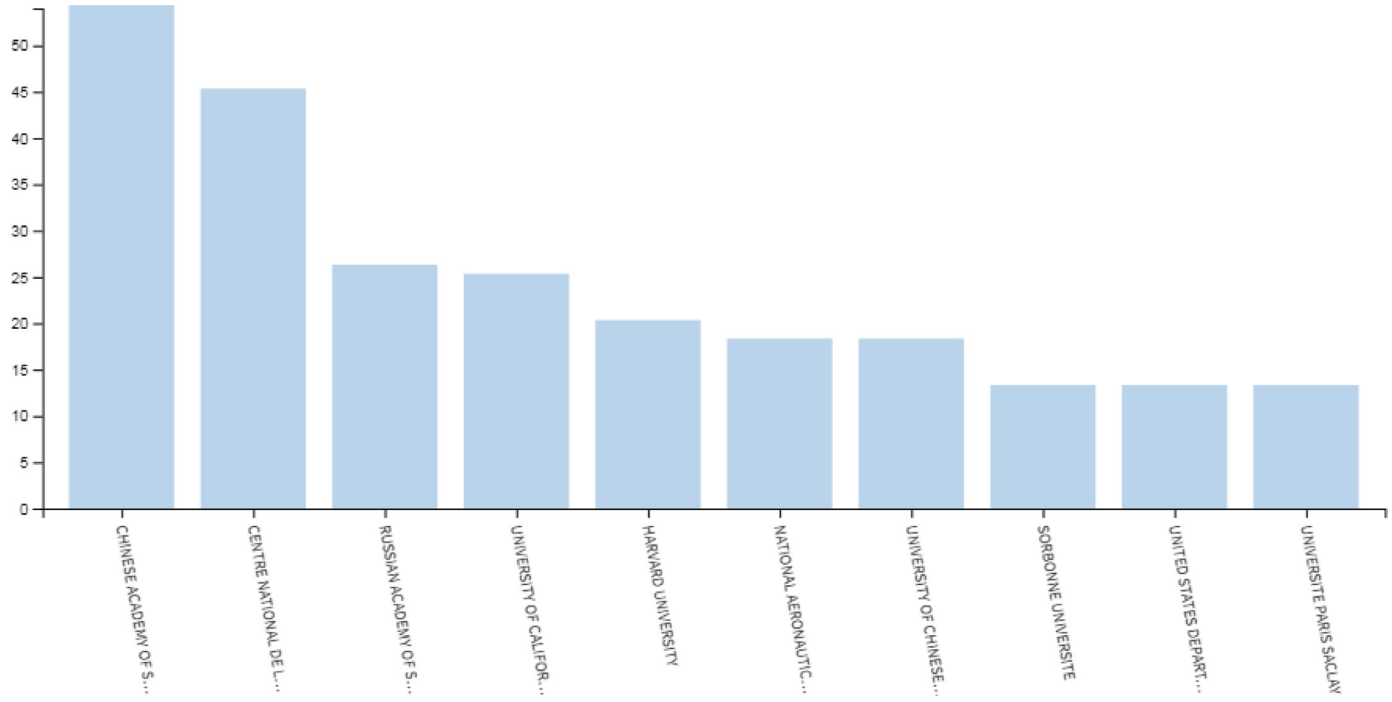

Fig. 7. Statistics of the best 10 institutions published about Covid-19 on Web of Science.

Choosing the right database, the right keywords and applying the filters that reflect the research objectives is a crucial step to have reliable results.

Among the credible scientific database which brings together most of the publishing houses known as Elsevier, Taylor & Francis, Springer..., we find Scopus, web of Science and for the medical field Pubmed [10] equipped with different filters to refine the search and limit the results found.

Some researches try to analyze data coming from the various scientific databases, but there are structural differences between the platforms. Thus the differences in the classification of information adopted by each of them builds an obstacle for an exploitation of the common data.

For a good bibliometric analysis, we choose the following bibliometric data:

- Article title.
- Authors.
- Keywords.
- Number of citations.
- Year of publication.
- Journals.
- Type of documents.
- Institution.
- Country.
- Field of research.

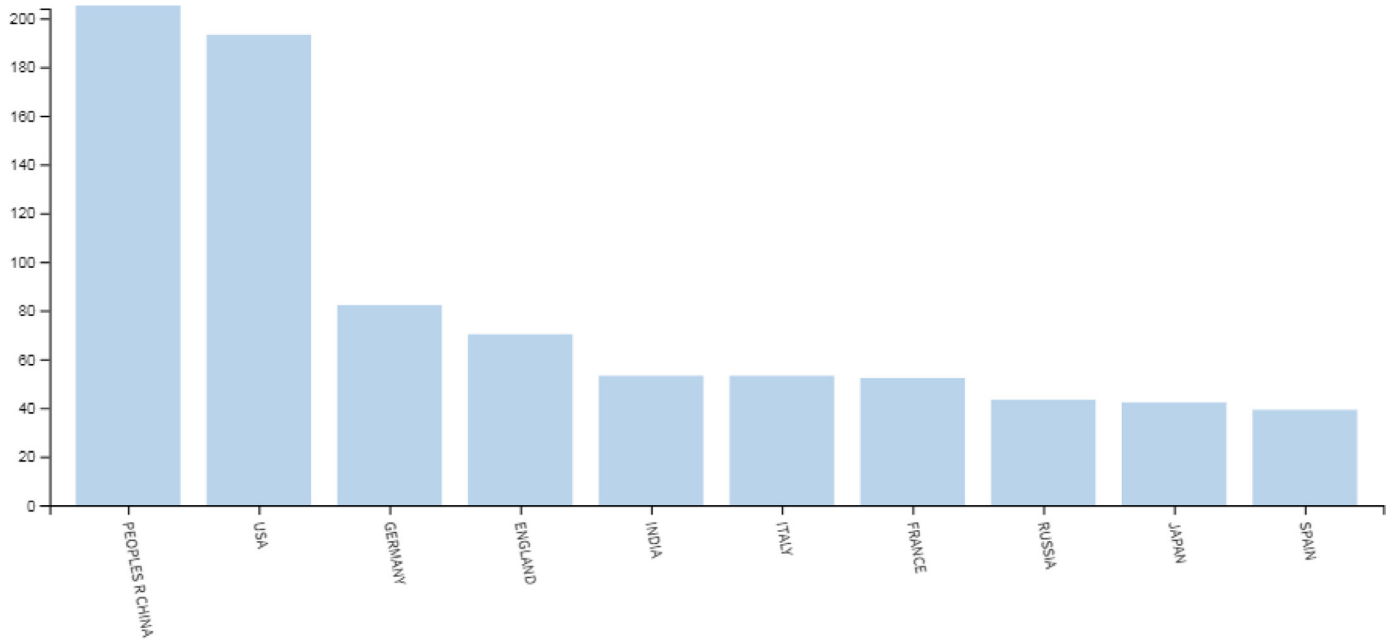

Fig. 8. Statistics of the best 10 countries published about Covid-19 on Web of Science.

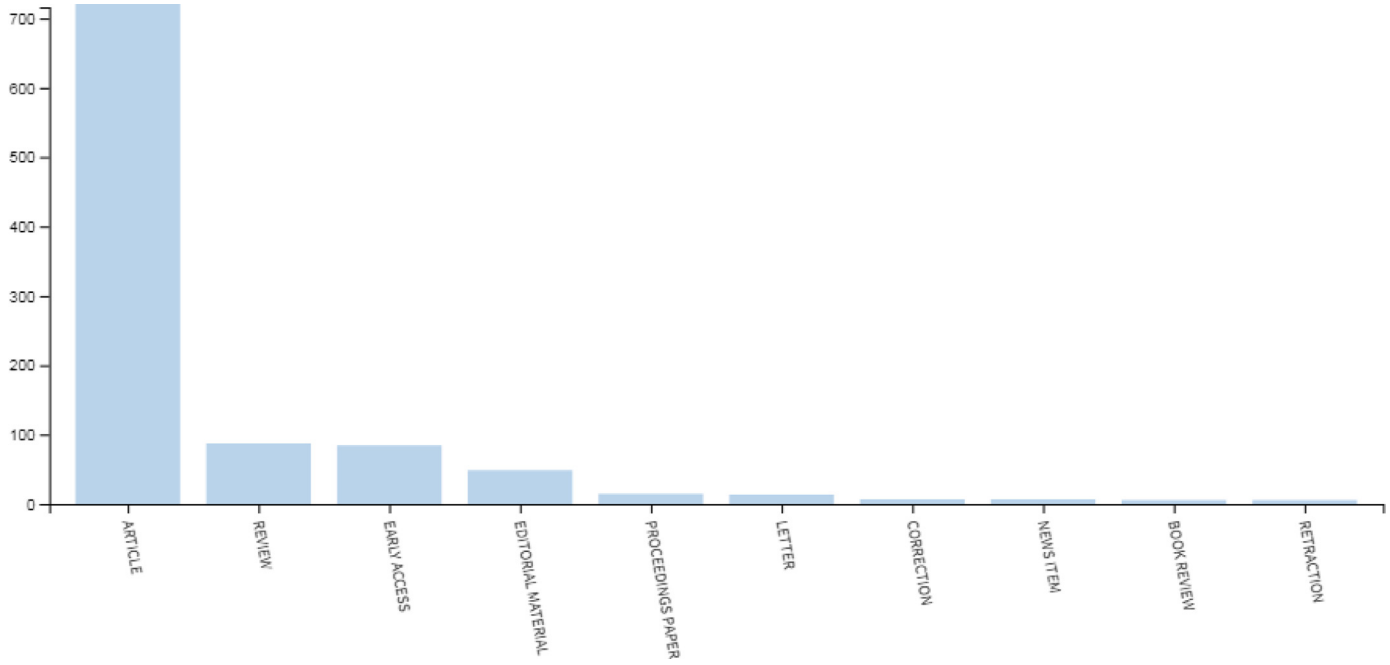

Fig. 9. Statistics of the type of document published about Covid-19 on Web of Science.

Regarding the indicators used by Scopus we find:

- H-index [11]: is based on the highest number of articles with at least the same number of citations.
- CiteScore: measures the average number of citations received per document published in the serial publication.
- SJR: measures the weighted citations received by the periodical, the weighting of the citations depends on the domain and the prestige of the citing series.
- SNIP: the standardized paper impact of the source which measures the actual citations received compared to the expected citations for the field of serial publication.

Regarding the indicators used by Web of Science we find:

- H-Index: the most used research indicator that measures both the productivity and the impact of an author's scientific production.
- The impact factor: measures the importance of a review according to the number of citations received in a year.
- Journal Citation Reports: Web of science product and an authoritative resource for impact factor data.

In the present case study, the keywords employed are "Covid-19" / "Coronavirus" from the beginning of 2020 (date of the start of the pandemic). The search should focus mainly on the

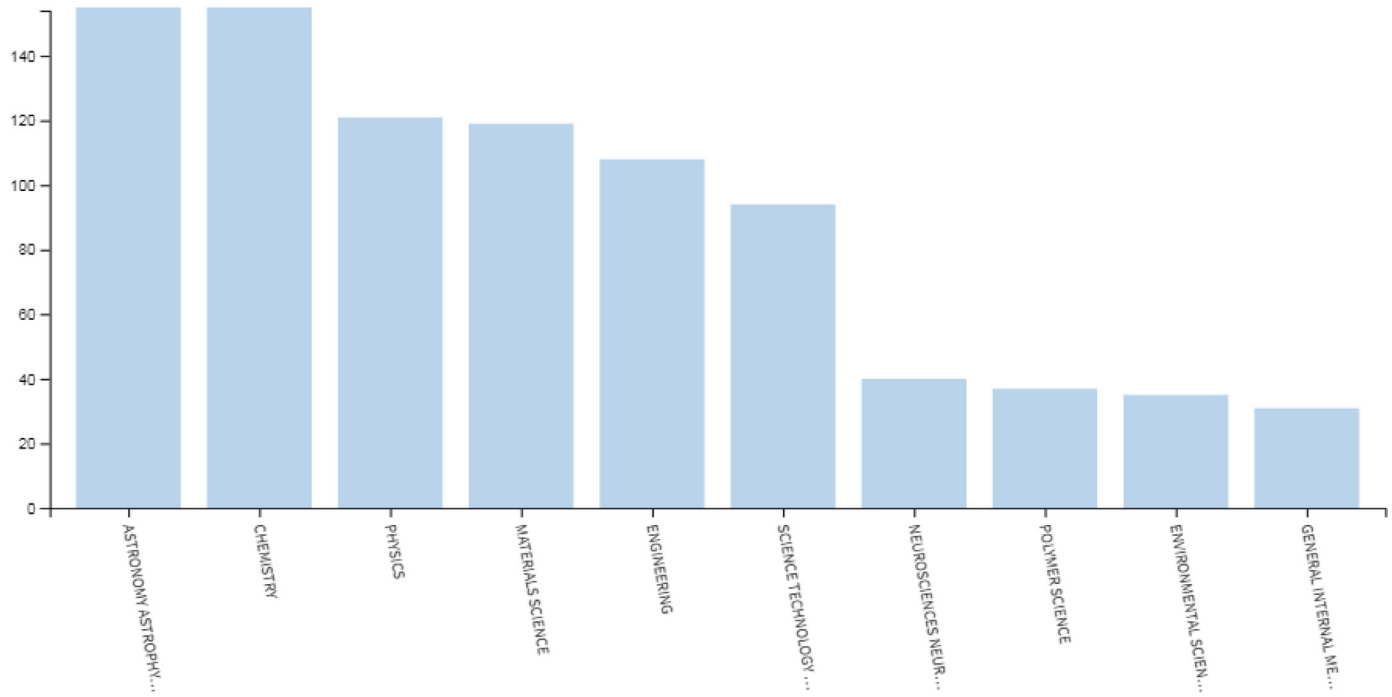

Fig. 10. Domain statistics published for Covid-19 on Web of Science.

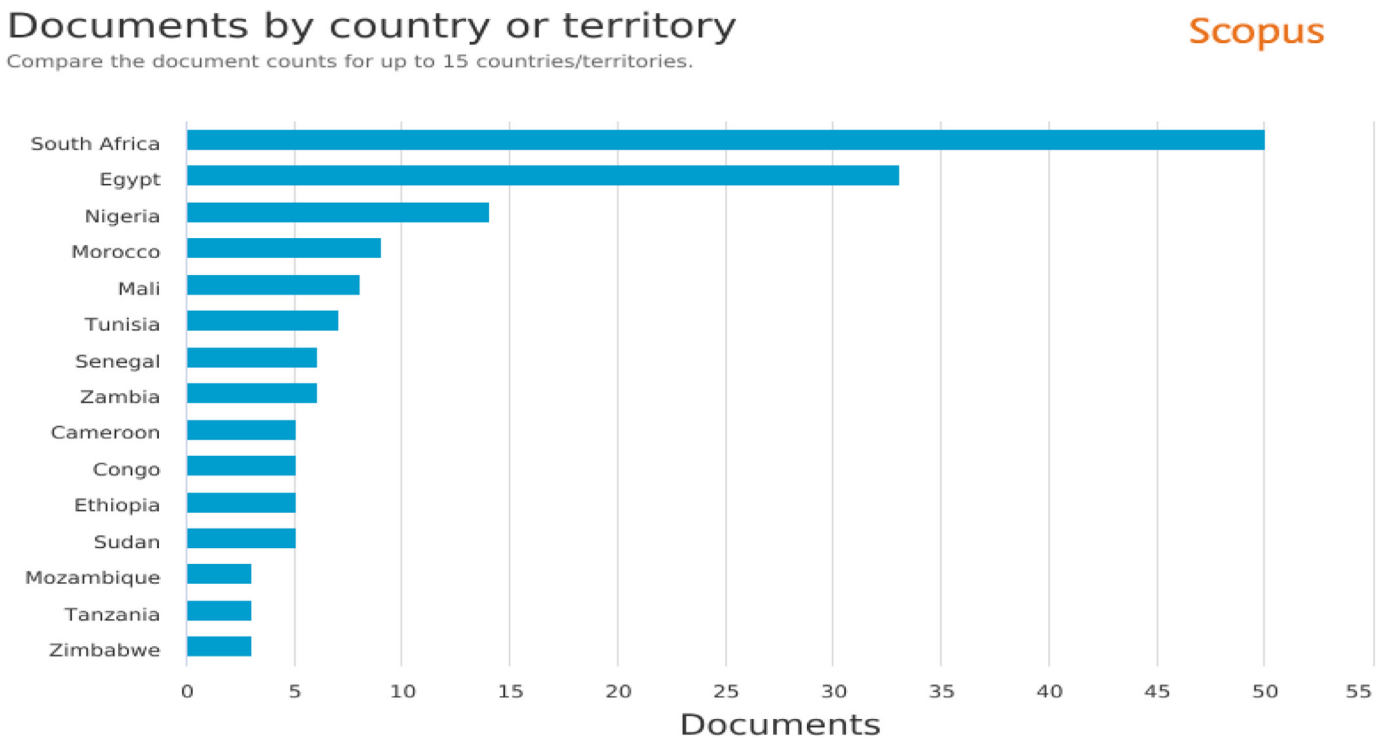

Fig. 11. Statistics of the best 10 African countries published for Covid-19 on Scopus.

titles, keywords and abstracts of articles in each of the databases. Then the results found for each of the three databases (Scopus, Web of science, Pubmed) builds our separate database on which our bibliometric analysis will be applied. We export the data from Scopus in format (.csv), Web of science, Pubmed in format (.txt).

Next, we use the VOSviewer software [12] which represents a high-performance solution with numerous viewing options with co-quotation, co-word, co-author network analysis.

4.1. Identification and analysis of research trends on Covid-19

Through bibliometric analyzes we try to get the trends of scientific research in the theme of Covid-19.

4.1.1. Analysis of authors, institutions and countries

In order to observe and evaluate the trends in publications in the thematic of Covid-19, the VOSviewer software was used to analyze the academic literature and examine the evolution of pub-

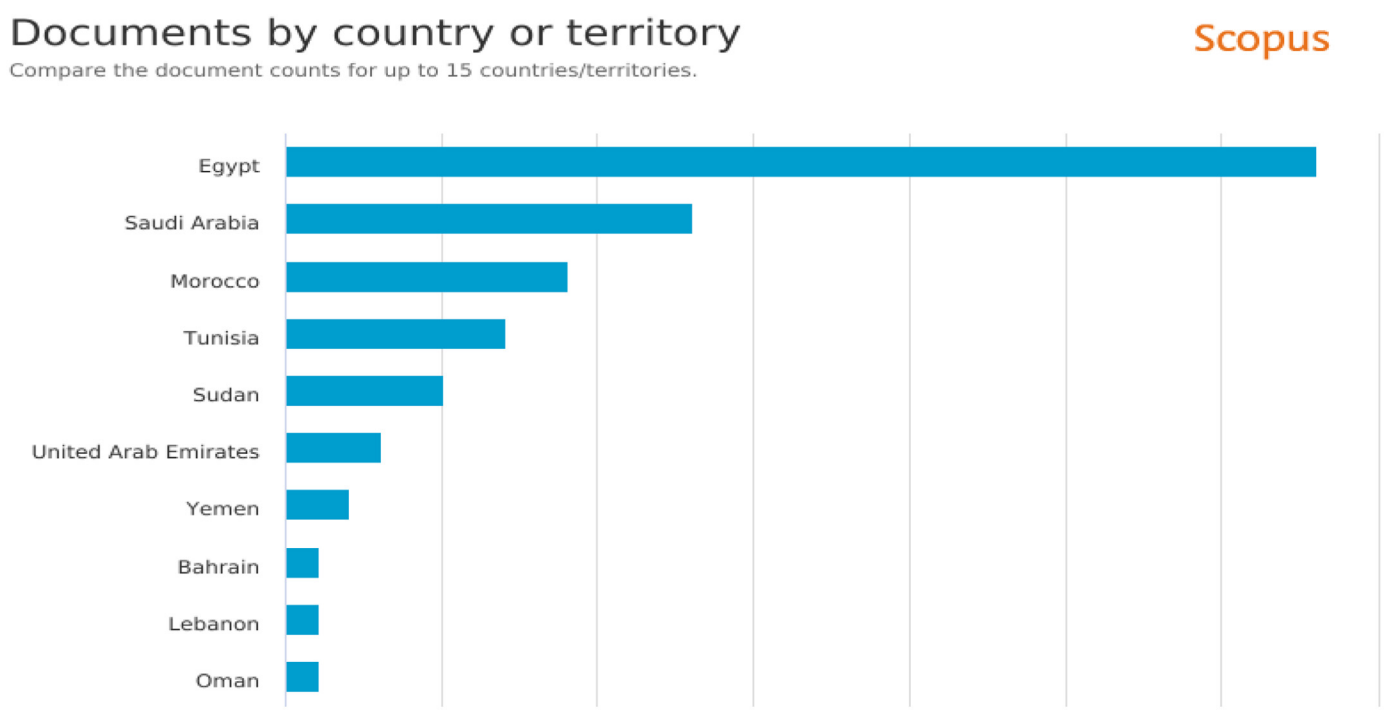

Fig. 12. Statistics of the best 10 Arab countries published about Covid-19 on Scopus.

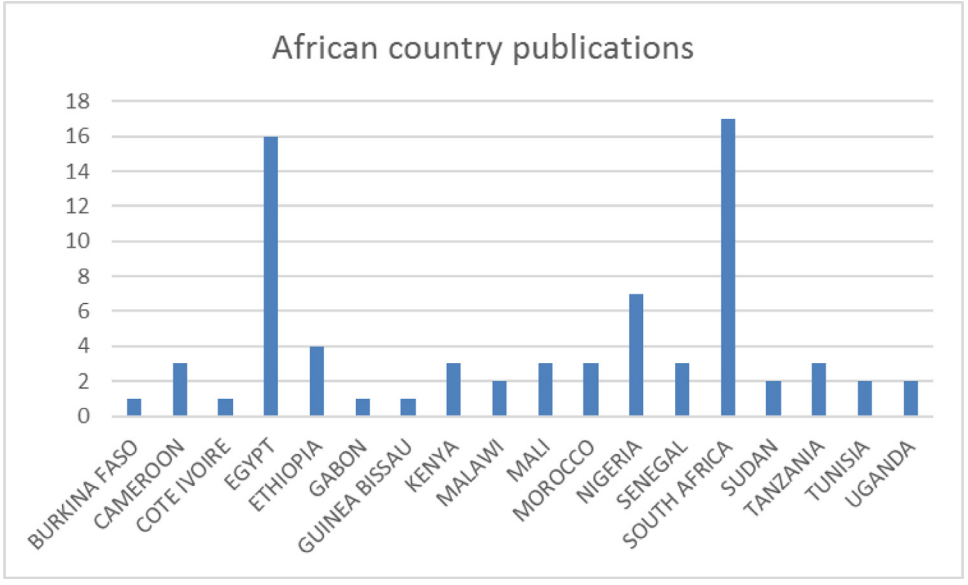

Fig. 13. Statistics of the best 10 African countries published about Covid-19 on Web of Science.

lished articles, co-authorship, geographic area (country) of authors, co-citation, co-occurrence.

The analysis of the authors belonging to the database allows to have a global view on the authors active in the thematic by offering the possibility to follow the work of these researchers by opening the door to achieve cooperation and partnerships.

Thus, the analyzes of research institutions and countries constitute an effective asset for finding the pillar institutions in each field, with the aim of seeking possible cooperation at the level of research institutions.

The software used for viewing and mapping the structure of a research are including Bibexcel, Histcite, Citespace, Gephi, and VOSviewer. For this work, we chose to work with VOSviewer because it allows us to easily display and interpret the display of large bibliometric maps.

In order to carry out the various analyzes previously cited and to examine the evolution of the articles published, we have for:

❖ Scopus:

- For authors:

We have 21 clusters distributed as follows:

Cluster 1-2: 42 items; Cluster 3: 29 items; Cluster 4-5: 27 items;
Cluster 6-7-8: 26 items; Cluster 9: 25 items; Cluster 10: 23 items;

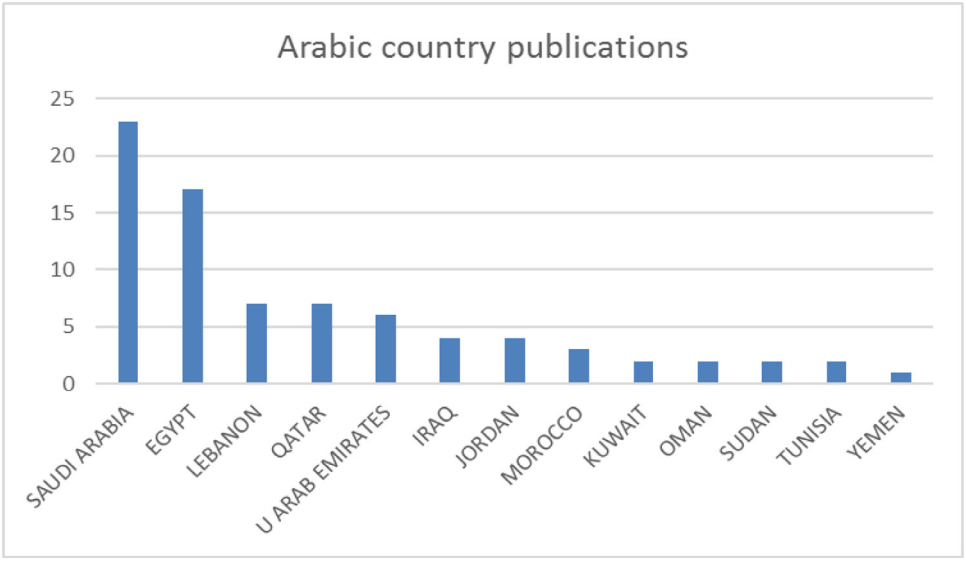

Fig. 14. Statistics of the best 10 Arab countries published about Covid-19 on Web of Science.

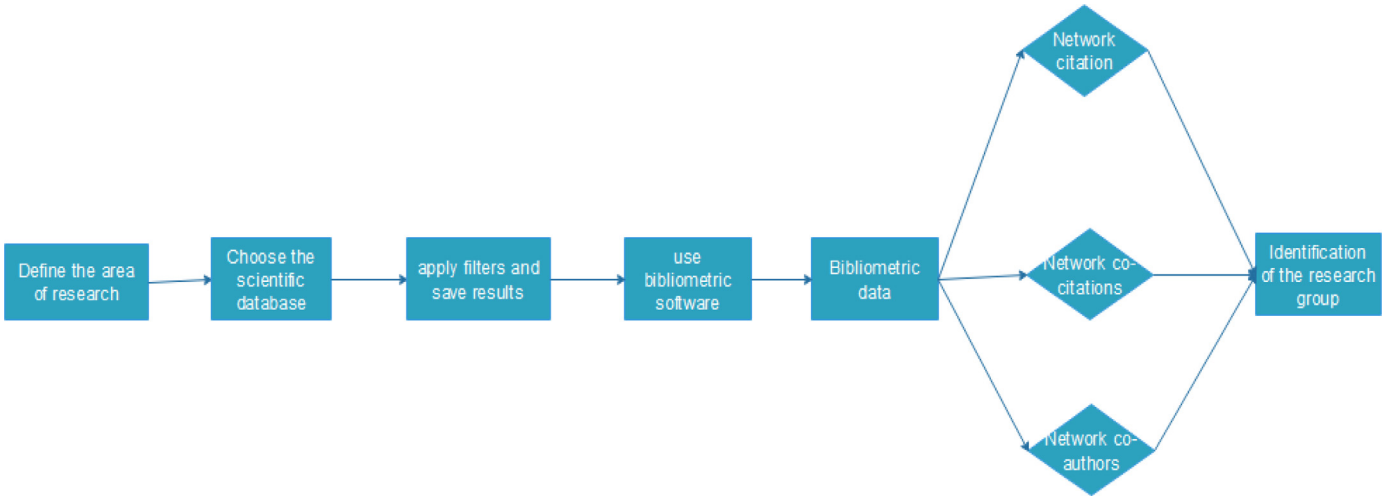

Fig. 15. Stage of the proposed bibliometric method.

Cluster 11: 21 items; Cluster 12-13: 19 items; Cluster 14-15: 16 items;
Cluster 16:15 items; Cluster 17: 14 items; Cluster 18: 13 items;
Cluster 19-20: 11 items; Cluster 21: 7 items.

The results clearly show that there are 21 groups of researchers collaborating with each other.

- For institutions:

We have 1 cluster which contains 12 items.
We deduce that most institutions collaborate with each other on an international scale and not at the regional or continental level.

- For countries:

We have 9 clusters distributed as follows:

Cluster 1-2-3: 5 items; Cluster 4-5-6: 4 items; Cluster 7-8-9: 3 items.

As we see in Fig. 21, the map indicates a large node representing China which means the great involvement of the Chinese giant through these researchers in the various research fields related to Covid-19.

❖ World of Science:

- For authors:

Bibliometric studies are used to identify networks of researchers or to map the structure of researchers in a given research area.
We have 9 clusters distributed as follows:

Cluster 1:46 items; Cluster2: 46 items; Cluster3: 20 items; Cluster 4:16 items;
Cluster 5:15 items; Cluster 6:11 items; Cluster 7: 11 items; Cluster 8: 10 items;
Cluster 9: 10 items.

The results clearly show that there are 9 groups of researchers who collaborate. Two groups have a significant number of researchers despite an exponential increase in the number of publications since the start of the pandemic, international collaboration between the authors remains low.

- For institutions:

The network analysis of research institutions with the highest number of links in this area are the institutions of the

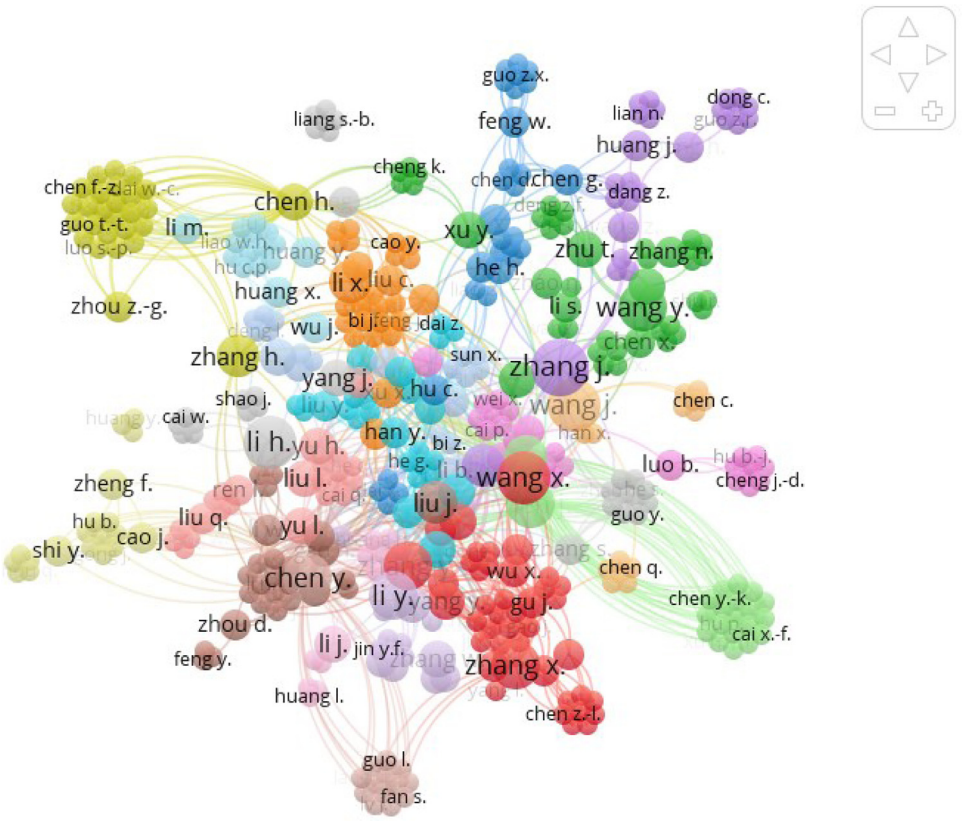

Fig. 16. Author co-authorship network in the "Network visualization" display mode.

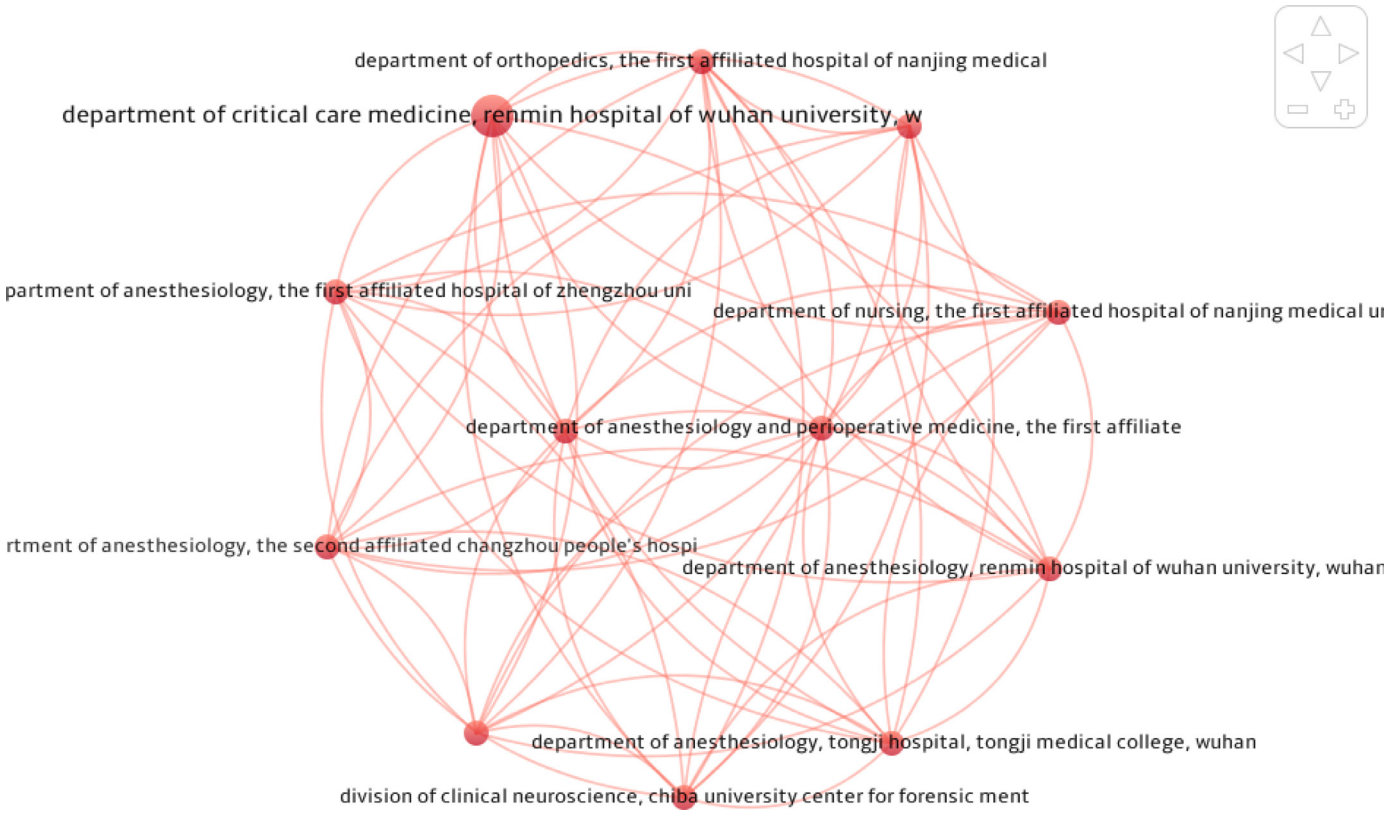

Fig. 17. Author organizations network in the "Network visualization" display mode.

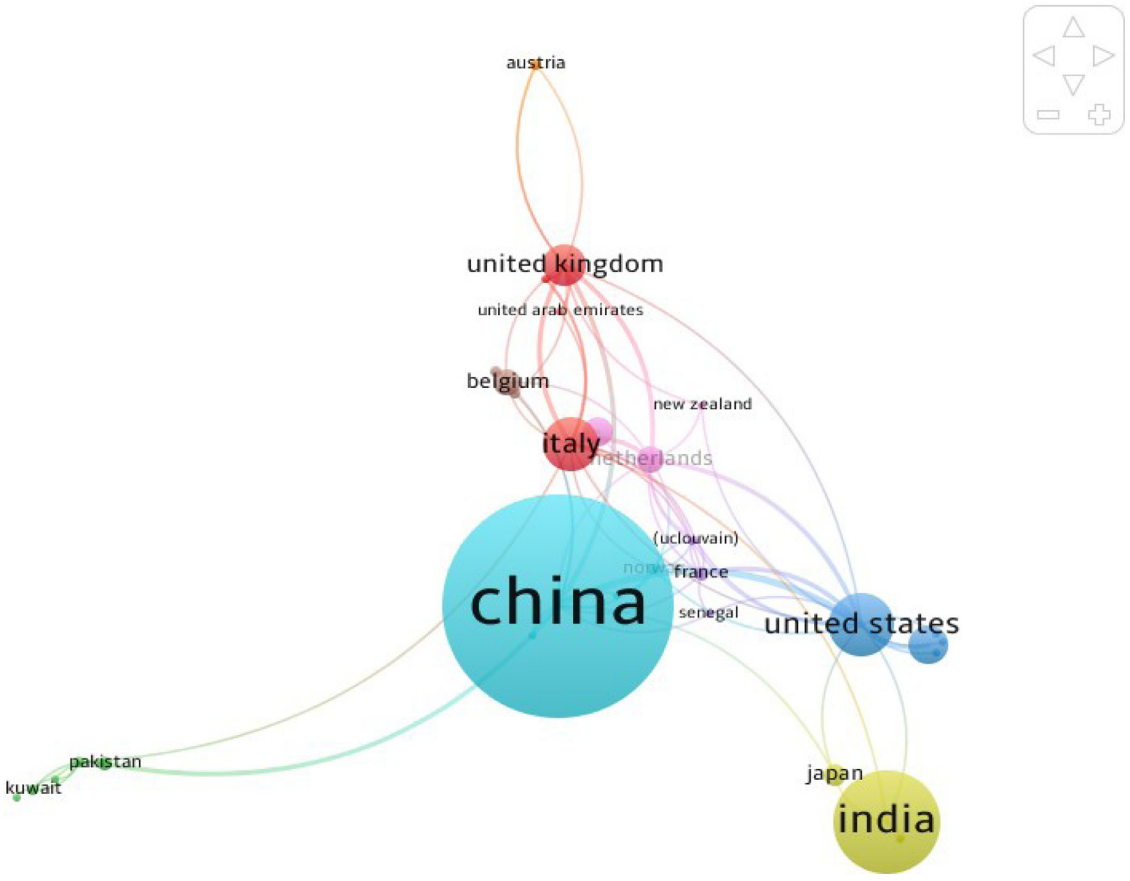

Fig. 18. Country organizations network in the "Network visualization" display mode.

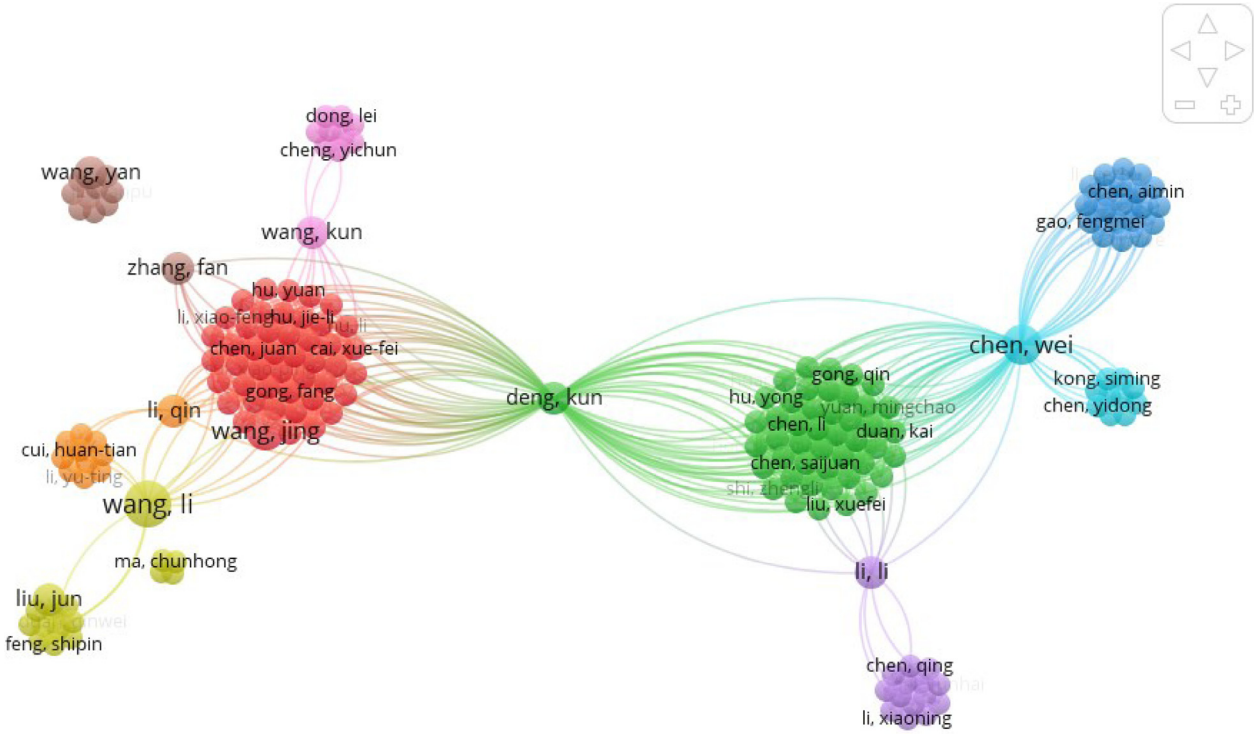

Fig. 19. Author co-authorship network in the "Network visualization" display mode.

Fig. 20. Author organizations network in the "Network visualization" display mode.

United States and China. In other words, the institutions in China and the United States have the highest total liaison force for collaboration with various institutions from different continents.

We have 31 clusters distributed as follows:

Cluster 1:33 items; Cluster2: 31 items; Cluster3: 30 items; Cluster 4:28 items;
Cluster 5-6: 27 items; Cluster 7-8: 25 items; Cluster 9-10: 24 items;
Cluster 11-12: 23 items; Cluster 13-14: 21 items; Cluster 15-16: 20 items;
Cluster 17-18: 16 items; Cluster 19-20: 15 items; Cluster 21: 14 items;
Cluster 22: 13 items; Cluster 23: 10 items; Cluster 24: 9 items;
Cluster 25-26: 7 items; Cluster 27: 6 items; Cluster 28-29-30-31: 5 items.

From the results found, it can be deduced that geographic proximity between institutions tends to strengthen the collaborative relationships of institutions. Thus, it warns of the need to expand cooperation in other regions, countries or continents.

- For countries:

The analysis of the network of countries is an important form of analysis which makes it possible to visualize the most influential countries in a given field of research, thus it exposes the degree of scientific cooperation between the countries.

We have 11 clusters distributed as follows:

Cluster 1: 7 items; Cluster 2-3: 6 items; Cluster 4: 5 items; Cluster 5: 4 items;
Cluster 6-7-8: 3 items; Cluster 9-10-11: 2 items.

As we can see in Fig. 18, the map shows a large node representing the countries and regions with the highest number of publications: China, United States, Italy, England, France and Spain Figs. 19 and 20.

❖ Pubmed:

- For authors:

We have 6 clusters distributed as follows:

Cluster 1:27 items; Cluster 2-3-4: 15 items; Cluster 5: 7 items; Cluster 6: 4 items.

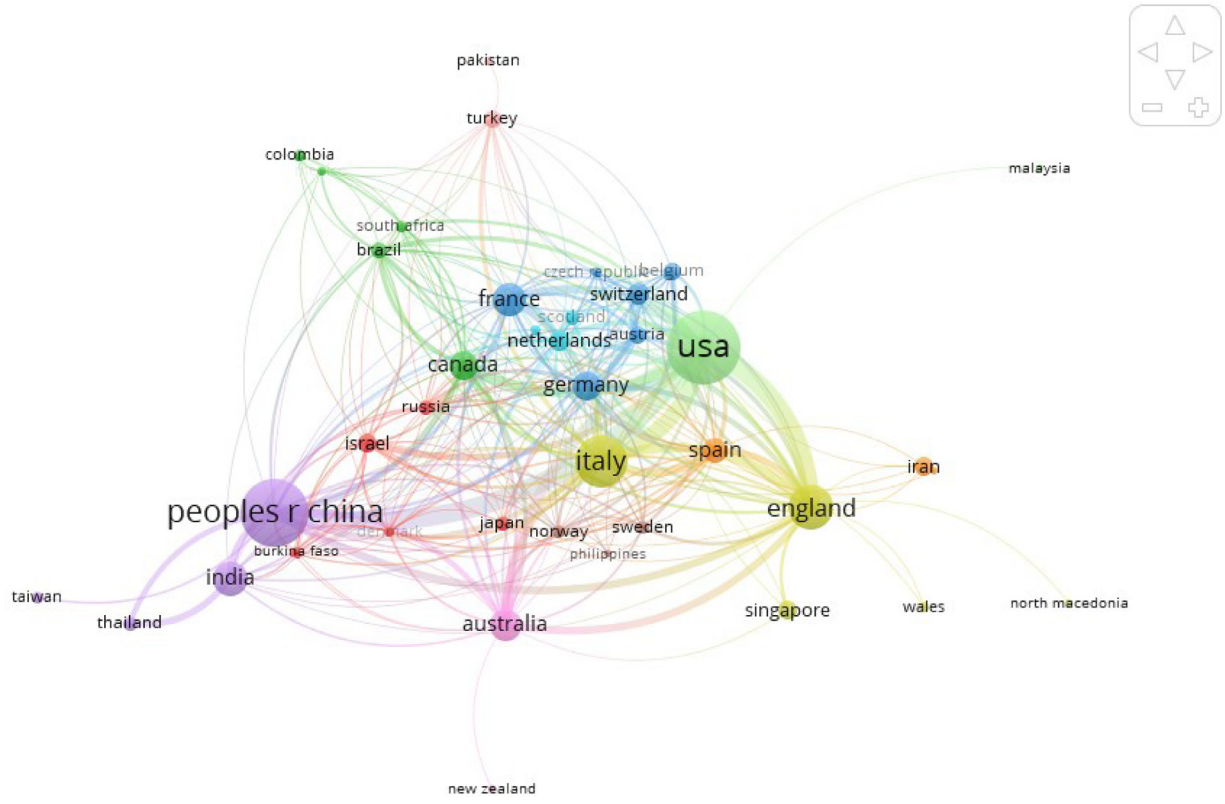

Fig. 21. Country organizations network in the "Network visualization" display mode.

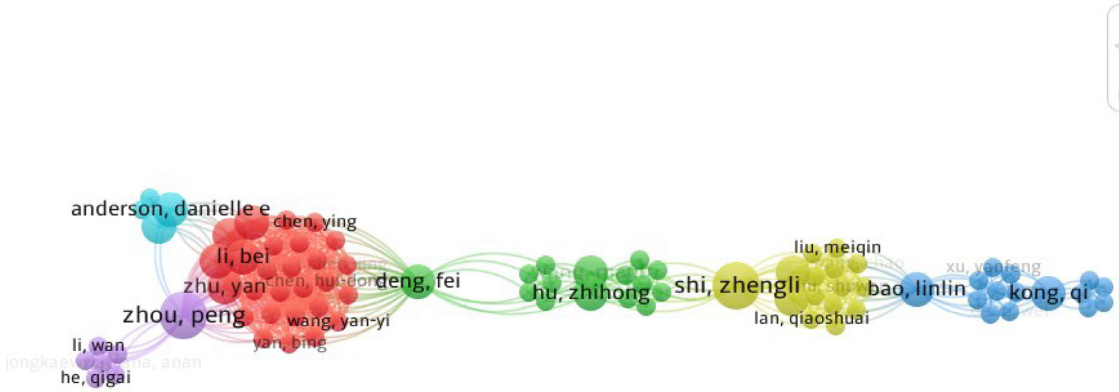

Fig. 22. Author co-authorship network in the "Network visualization" display mode.

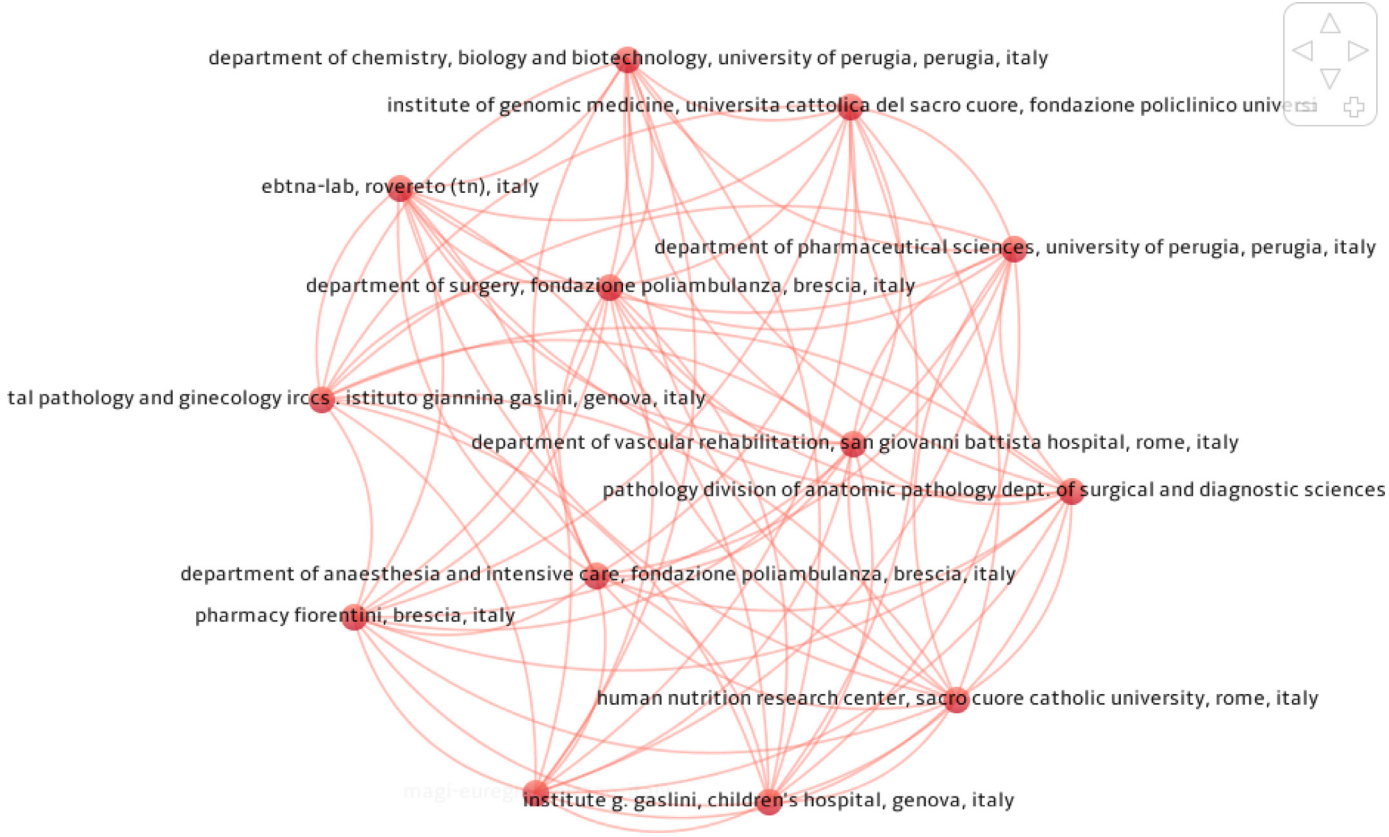

Fig. 23. Author organizations network in the "Network visualization" display mode.

The results clearly show that there are 6 groups of researchers who collaborate with each other, a group has a large number of researchers, followed by a group that is distinguished by the number of researchers who compose them.

- For institutions:

In 1 cluster with 13 items, we notice that there is a significant presence of Italian medical institutions, the analysis of data from Pubmed by VOSviewer does not offer the possibility of analyzing the network of countries.

4.1.2. Analysis of keywords

❖ VOSviewer:

We have 3 clusters distributed as follows:

Cluster 1: 6 items; cluster 2-3: 4 items.

The results found build a map dividing the keywords into three groups with the minimum number of occurrences of a keyword fixed at 6 elements for the first group and 4 elements for the second and third group. The keyword "Coronavirus" has the highest occurrence and total binding strength, other keywords with a high occurrence include "Sars-cov-2", "Covid-19" Figs. 22 and 23.

❖ Wordle [13]:

Among the existing display means, there is the word cloud which is a practical tool allowing to have a dimensional visualization of the keywords most used in the database. For our case, we use wordle which is an analysis tool which makes it possible to display a word cloud which gives greater importance to the words which appear more frequently in the source text, for the three scientific databases already mentioned, we find:

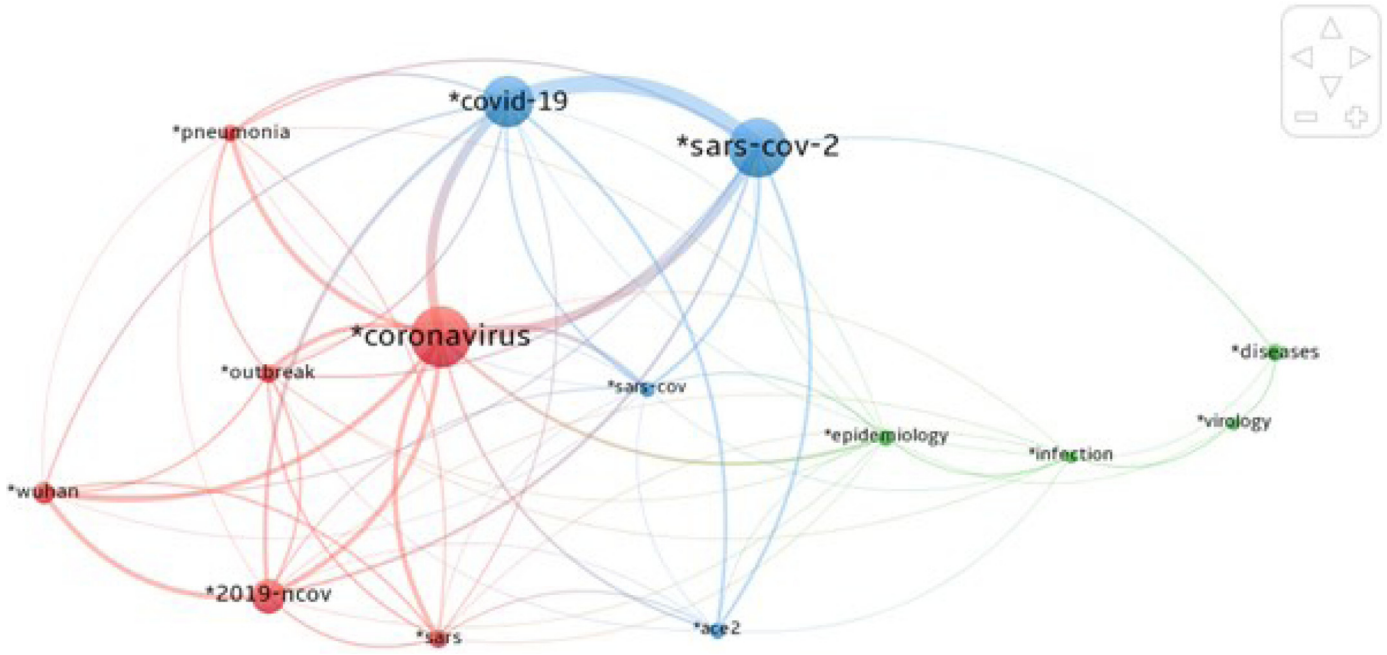

Fig. 24. Author keywords network in the "Network visualization" display mode.

Fig. 25. Wordle-generated image of the words Covid-19 for Scopus.

Fig. 26. Wordle-generated image of the words Covid-19 for Web of science.

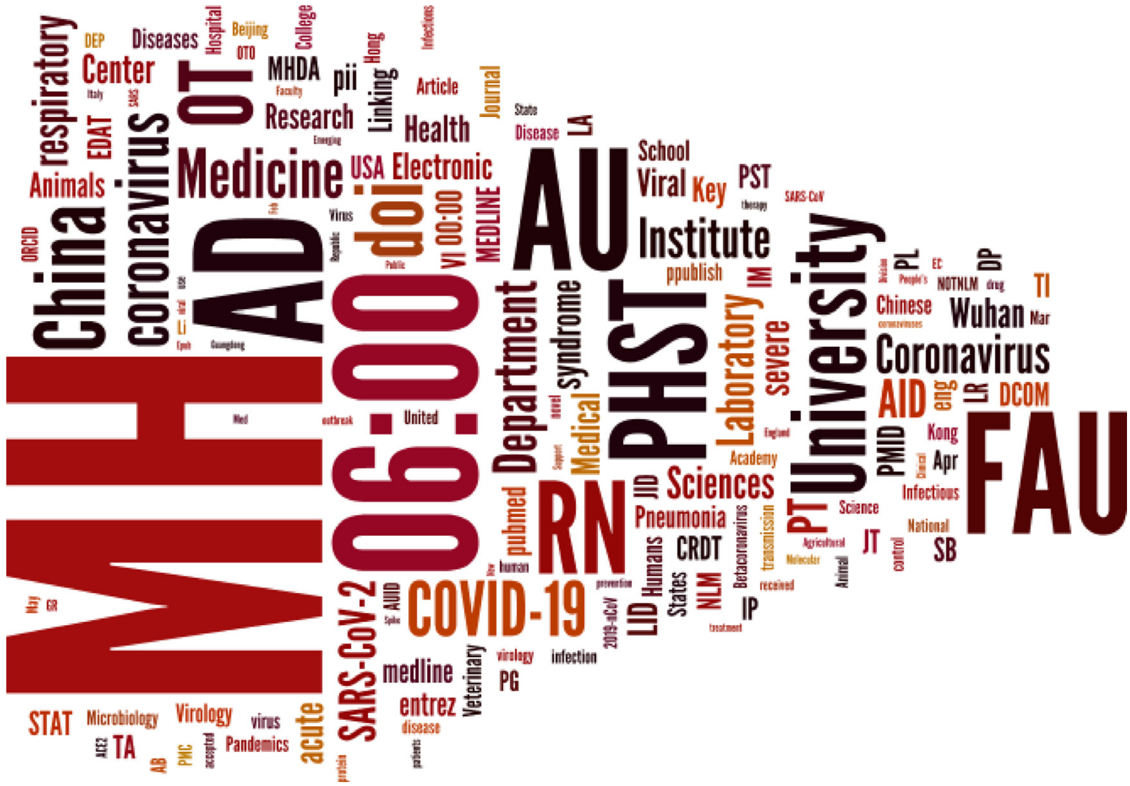

Fig. 27. Wordle-generated image of the words Covid-19 for Pubmed.

- Scopus:
- Web of Science:
- Pubmed:

From the Figs. 24–27 provided by VOSviewer and Wordle, a set of words related to the pandemic such as (Covid-19, Coronavirus, Sars-cov-2, 2019-ncon) as synonyms used in scientific literature, so the appearance of terms (China, Wuhan, USA) refers to the place of appearance of the pandemic and the countries that are conducting research to find the vaccine, too (Medical, Health, Hospital, virology) refers to the most concerned research area, (Zhang, Wang) for the most productive researchers in the topic of Covid-19 in scientific databases.

5. Conclusion

Since the onset of the pandemic, considerable effort has been invested by researchers worldwide depending on the fields and resources available, an exponential increase in scientific production has been recorded in the various databases around the Covid-19.

In this work, we opted for a statistical study for the data from the bibliographic databases Scopus, Web of Science for the theme of Covid-19. The scientific contribution of researchers from the USA and China shows a total involvement of institutions from these two countries, so for the African continent researchers from "South Africa and Egypt are the exception, while for the Arab region Saudi Arabia and Egypt are leading the efforts of the Arab countries for this pandemic.

Afterwards, a bibliometric analysis method was adopted in order to map the state of the art on the theme of Covid-19, so the three scientific databases (Scopus, Web of Science, Pubmed) were used. Thus, the search must be precise and planned by combining the precision of the terms to be used and adequate filters to refine the results found, in order to conduct a relevant bibliometric analysis by analyzing the contributions of the authors, institutions, countries and the words- keys.

Finally, it is well known that the method presented remains applicable for other scientific themes and not only for the Covid-19 theme, it should be noted that the results obtained with the application of the proposed method may vary depending on the basis of scientific data chosen and the appropriate filters in order to present the evolution of published articles, co-authors, geographic area of the authors, co-citation, co-occurrence analysis and keywords.

Declaration of Competing Interest

We wish to draw the attention of the Editor to the following facts which may be considered as potential conflicts of interest and to significant financial contributions to this work. [OR] We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome. We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us. We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

References

[1] Sengupta IN. Bibliometrics, informetrics, scientometrics and librametrics: an overview. Libri 2009;42(2):75–98.
[2] Tiernoa NR, Gonzalez-Cruz TF, Martinez JL. An overview of qualitative comparative analysis: a bibliometric analysis. J Innov Knowl 2017;2(1):15–23.

[3] Mingers J, Leydesdorff L. A review of theory and practice in scientometrics. Eur J Oper Res 2015;246(1):1–19.
[4] Kelleher C, Wagener T. Ten guidelines for effective data visualization in scientific publications. Environ Model Softw 2011;26(6):822–7.
[5] https://coronavirus.jhu.edu Accessed 23/05/2020.
[6] https://www.who.int Accessed 23/05/2020.
[7] https://www.scopus.com Accessed 23/05/2020.
[8] https://webofknowledge.com Accessed 23/05/2020.
[9] M. Franceschet, "A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar", Vol 83(1), Pages 243–258, 2010.
[10] https://pubmed.ncbi.nlm.nih.gov Accessed 23/05/2020.
[11] Bornmann L, Daniel H-D. Does the h-index for ranking of scientists really work? Scientometrics 2005;65:391–2.
[12] van Eck N, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics 2010;84(2):523–38.
[13] Viegas FB, Wattenberg M, Feinberg J. Participatory visualization with Wordle. IEEE Trans Vis Comput Graph 2009;15(6):1137–44.