République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Akli Mohand Oulhadj de Bouira

Faculté des Sciences et des Sciences Appliquées

Département d'Informatique

Mémoire de Master

en Informatique

Spécialité : *Isil & Gsi*

Thème

Regroupement des documents scientifiques basé sur la similarité des citations et du texte.

Encadré par

— M. BAL Kamel.

Réalisé par

— Mlle Taoui Basma

— Mlle Zeggane Hind

2022/2023

Autorisation

الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire

وزارة التعليم العالي و البحث العلمي
جامعة أكلي محند أولحاج - البويرة

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Akli Mohand Oulhadj - Bouira
Tasdawit Akli Muhend Ulhag - Tubirett

التصريح الشرفي الخاص بالالتزام بقواعد النزاهة العلمية لإنجاز بحث

انا الممضي اسفله،

السيد(ة) طاوي بسمة الصفة طالب (ماستر / دكتوراه)

الحامل(ة) لبطاقة التعريف الوطنية .... والصادرة بتاريخ ....

المسجل(ة) بكلية / معهد العلوم والعلوم التطبيقية قسم إعلام آلي

تخصص هندسة نظم المعلومات والبرمجيات

والمكلف(ة) بإنجاز اعمال بحث(مذكرة. التخرج. مذكره ماستر. مذكرة ماجستير. اطروحة دكتوراه).

عنوانها Scientific documents clustering based on citation and text similarity

أصرح بشرفي اني ألتزم بمراعاة المعايير العلمية والمنهجية الاخلاقيات المهنية والنزاهة الاكاديمية المطلوبة في انجاز البحث المذكور أعلاه

التاريخ: 24/06/2023

توقيع المعني (ة)

هيئة مراقبة السرقة العلمية:

النسبة: 21 %

البويرة في 2023/07/01

الامضاء

وزارة التعليم العالي و البحث العلمي
جامعة أكلي محند أولحاج - البويرة

التصريح الشرفي الخاص بالالتزام بقواعد النزاهة العلمية

لإنجاز بحث

انا الممضي اسفله.

السيد(ة) زقان سهند الصفة طالب (ماستر / دكتوراه)

الحامل(ة) لبطاقة التعريف الوطنية 403066593 والصادرة بتاريخ 28/09/2022

المسجل(ة) بكلية / معهد العلوم والعلوم التطبيقية قسم الاعلام الآلي

تخصص عسقدية أنظمة المعلوماتية

والمكلف(ة) بإنجاز اعمال بحث(مذكرة. التخرج. مذكرة ماستر. مذكرة ماجستير. اطروحة دكتوراه)

عنوانها Scientific documents clustering based on citation and text similarity

أصرح بشرفي اني ألتزم بمراعاة المعايير العلمية والمنهجية الاخلاقيات المهنية والنزاهة الاكاديمية المطلوبة في انجاز البحث المذكور أعلاه

التاريخ: 24/06/2023 توقيع المعني (ة)

هيئة مراقبة السرقة العلمية:

البويرة في: 2023/07/01

النسبة: 22 %

الامضاء

رئيس قسم العلوم الدقيقة
جوابري عبد الرزاق

Remerciements

Avant tout nous remercions dieu le tout puissant qui nous a donné la force, la patience et le courage pour qu'on puisse accomplir ce travail.

Nous remercions profondément notre encadreur monsieur **M.BAL Kamel**, pour ses suivis et ses précieuses orientations dans notre travail et Nous voudraient le remercier pour tous ses conseils et ses remarques intéressantes.

Nous exprimons nos reconnaissances à tous personnes qui a contribué de prés à l'achévement de ce travail ; nos enseignants, nos amis, nos collégues de promotion Isil et Gsi 2023 .

Nous remercions également les membres de jury d'avoir acceptée d'évaluer notre travail. Ainsi que tous les enseignants qui ont assuré notre formation en cursus universitaire.

Finalement, nous aimerions aussi remercier nos familles nos parents pour leur amour, leurs conseils ainsi que leur soutien inconditionnel, à la fois moral et économique, qui nous a permis de réaliser les études que nous voulont et par conséquent ce mémoire .

Dédicaces

Je dédie ce mémoire :

À mon pére qui a été mon guide , son soutien son affection et la confiance qu'il m'a accordé m'ont donné la motivation nécessaire pour aller au-delà de mes limites et atteindre mes objectifs.

À ma mère pour son amour ses encouragements et ses sacrifices dévoués ont été les piliers de ma vie.

À mon grand pére décédé , qui m'a toujours poussé et motivé dans ma vie particuliérement dans mes études puisse dieu le tout puissant l'avoir en sa sainte miséricorde.

À mes chéres soeurs " **RANIA** et **DINA**" et mes fréres" **YANI** et **YANIS**" et toute ma famille "**TAOUI** ".

À tous mes chers amies notamment mon binome "**ZEGGANE Hind**" , "**NADIR Sonia**" , "**DAHMOUNI Yasmine**" et à tous ceux qui m'aiment.

TAOUI Basma.

Dédicaces

Je dédie ce mémoire :

À ma mère, celle qui a illuminé ma vie de son amour inconditionnel et de sa bienveillance infinie.Sa présence dans ma vie est un cadeau précieux que je chéris chaque jour.

À mon père, qui a été mon guide et mon soutien tout au long de mon parcours. Sa confiance en moi et sa croyance en mes capacités m'ont donné la détermination nécessaire pour surmonter les obstacles et poursuivre mes rêves.

À mon cher jumeau"**WALID**", qui a été mon compagnon de voyage depuis le premier jour , pour son tendresse et son motivation et à mon petit frére "**SALAH**".

À tous mes chers amies notamment mon binome "**TAOUI BASMA**".

À tous les membres précieux de ma famille et à tous ceux qui m'entourent avec amour et soutien.

ZEGGANE Hind.

ملخص

البحث عن المعلومات في مجموعات الوثائق العلمية أصبح مهمة معقدة بشكل متزايد بسبب الكمية المتزايدة من المعلومات التي تنتج كل عام وتنوع المصادر المتاحة. يجب على الباحثين أن يقضوا ساعات عديدة في البحث عن المقالات ذات الصلة وقراءة وتحليل المعلومات، وهو عمل يستغرق وقتًا طويلاً ومرهق.

لتسهيل عملية البحث عن المعلومات العلمية، تم تطوير العديد من التقنيات والنهج. أحد هذه النهج هو تجميع الوثائق، والذي يتمثل في تجميع المقالات المماثلة في مجموعات أو عناقيد استنادًا إلى خصائصها المشتركة. يتيح هذا للباحثين تصفح مجموعات الوثائق بسهولة واكتشاف المقالات ذات الصلة بسرعة في مجال اهتمامهم.

في هذا السياق، يركز عملنا على تطوير طريقة لتجميع الوثائق العلمية بناءً على التشابه في الاقتباسات والنص. يتم استخدام خوارزمية كمانس لأداء هذا التجميع عن طريق استغلال المصطلحات من أقسام العنوان والملخص لحساب التشابه النصي، ومعلومات المراجع الببليوجرافية لحساب تشابه الاقتباسات لكل وثيقة. في النهاية، قمنا بتقييم عملنا باستخدام معامل السيلويت لقياس جودة التجميع.

الكلمات الرئيسية : المقال العلمي، التصنيف التلقائي، تجميع الوثائق النصية، شبكة الاقتباسات، التجميع، كمانس.

Abstract

The search for information within collections of scientific documents has become an increasingly complex task due to the growing amount of information produced each year and the diversity of available sources. Researchers have to spend many hours searching for relevant articles, reading and analyzing information, which can be a laborious and time-consuming process.

To facilitate this search for scientific information, numerous techniques and approaches have been developed. One of these approaches is document clustering, which involves grouping similar articles into clusters based on their common characteristics. This allows researchers to navigate through document collections more easily and quickly discover relevant articles in their field of interest.

In this context, our work focuses on developing a method for clustering scientific documents based on the similarity of citations and text. The K-means algorithm is used

to perform this clustering by leveraging the terms from the title and abstract sections to calculate textual similarity, and the information from the bibliographic references to calculate citation similarity for each document. Finally, we evaluated our work using the silhouette coefficient to measure the quality of the clustering.

Key words : Scientific article, automatic classification, Text document clustering, Citation network, Clustering, K-means.

Résumé

La recherche d'informations dans les collections de documents scientifiques est devenue une tâche de plus en plus complexe en raison de la quantité croissante d'informations produites chaque année et de la diversité des sources disponibles. Les chercheurs doivent consacrer de nombreuses heures à la recherche d'articles pertinents, à la lecture et à l'analyse des informations, ce qui peut être un processus laborieux et couteux en temps.

Pour faciliter cette recherche d'informations scientifiques, de nombreuses techniques et approches ont été développées. L'une de ces approches est le regroupement de documents, qui consiste à rassembler des articles similaires en groupes ou clusters en fonction de leurs caractéristiques communes. Cela permet aux chercheurs de naviguer plus facilement dans les collections de documents et de découvrir rapidement des articles pertinents dans leur domaine d'intérêt.

Dans ce contexte, notre travail se concentre sur le développement d'une méthode de regroupement de documents scientifiques basée sur la similarité des citations et du texte. L'algorithme K-means est utilisé pour effectuer ce regroupement en exploitant les termes des sections titre et résumé pour calculer la similarité textuelle et les informations des références bibliographiques pour calculer la similarité des citations de chaque document. Finalement nous avons évalué notre travail en utilisant le coefficient de silhouette pour mesurer la qualité du Clustering.

Mots clés : Article scientifique, classification automatique, Clustering de document textuels, Réseau de citations , Clustering, K-means.

Table des matières

Table des figures

Liste des tableaux

Introduction générale

1. Contexte et problématique

Le domaine de la recherche scientifique est en constante expansion, avec une quantité impressionnante de documents publiés chaque année dans diverses disciplines. Cette abondance d'informations rend essentiel le développement de méthodes efficaces pour organiser, structurer et extraire des connaissances à partir de ces grandes sources de documents.

Les documents scientifiques représentent un moyen crucial de communication et de diffusion des connaissances dans la communauté scientifique. Ils englobent des articles de recherche, des revues, des thèses, des rapports techniques et bien d'autres formes de documentation qui contribuent à l'avancement des connaissances et à la compréhension des divers domaines scientifiques.

Cependant, avec cette expansion constante, il devient de plus en plus difficile pour les chercheurs et les praticiens de gérer et d'exploiter efficacement cette masse de documents scientifiques. La recherche d'informations pertinentes, la détection de problématiques de recherche émergentes et l'identification de liens entre les travaux deviennent des tâches complexes.

Cette situation soulève la nécessité d'utiliser des techniques pour organiser et structurer les collections de documents scientifiques de manière à faciliter leur exploration et leur utilisation. Le clustering, également connu sous le nom de regroupement ou de classification non supervisée, se place comme une approche prometteuse pour répondre à ce défi. Il permet de regrouper automatiquement les documents qui partagent des similitudes, ce qui facilite leur exploration et leur organisation. Il peut révéler des tendances, des thématiques émergentes et des relations entre les travaux de recherche, permettant aux chercheurs de découvrir de nouvelles perspectives et d'identifier des documents pertinents

dans leur domaine d'intérêt.

Les méthodes de clustering traditionnelles utilisent souvent des mesures de similarité ou de distance pour évaluer la proximité entre les documents. Cependant, le clustering des documents scientifiques basé uniquement sur des mesures de similarité a base du contenu textuel seulement peut être limité, car il ne prend pas en compte la richesse et la spécificité des documents scientifiques comme : la structure, les métadonnées (auteurs, mots-clés, nom de revue, ...etc) et surtout les informations de citation. Le document scientifique n'est pas un texte brut seulement, mais il répond à des exigences qui le rends très riches en termes d'information supplémentaires qui peuvent être exploitées durant la phase de représentation et de clustering de ces documents.

Les articles scientifiques ont des liens de références et de citations, créant ainsi un réseau de connaissances interconnectées. Les citations permettent de découvrir des relations entre les articles, d'identifier les documents influents et de suivre l'évolution des connaissances. Intégrer ces informations de citation dans le processus de clustering peut améliorer la pertinence des regroupements obtenus et faciliter la navigation dans les collections d'articles scientifiques.

2. Objectifs

Nous envisageons à travers ce travail de développer un système de Clustering des articles scientifiques en combinant les mesures de similarité à base du contenu textuel des articles et des informations de bibliographie. En intégrant à la fois la similarité textuelle et les liens de citation dans le processus de regroupement, nous visons à créer des clusters plus cohérents et plus informatifs. La similarité à base de contenu textuel permettra de capturer les similitudes thématiques entre les documents et la similarité à base de citation permettra de capturer les relations entre les travaux basées sur les références. En exploitant ces deux sources d'informations complémentaires, nous espérons obtenir des regroupements plus pertinents et faciliter l'accès aux connaissances scientifiques.

Dans ce cadre nous envisageons d'évaluer plusieurs variantes (contenu seulement, contenu + citations) pour évaluer l'impact de la prise en compte des différentes composantes de l'articles scientifique : Titre, résumé, références dans la qualité du clustering.

3. Organisation du mémoire

Ce mémoire est organisé en quatre chapitres :

Chapitre 1 : Généralités sur les documents scientifiques : dans ce chapitre

nous définissons le domaine de notre etude qu'est le document scientifique , leurs caractéristiques, leurs types et leur structure.Comme nous touchons les citations et leurs éléments essentiels.

Chapitre 2 :Clasification automatique de documents : ce chapitre présente unétat de l'art sur l'apprentissage automatiquenous et le clustering commence par aborder le domaine de l'apprentissage automatique (Machine Learning). ensuite présente la technique du Clustering son principe de fonctionnement, ces algorithmes ainsi que les mesures de son évaluation.Enfin il abordele Clustering des documents textuels et ses spécificités.

Chapitre 3 :Système de Clustering des documents scientifiques :c'est la partie de conception qui permet de défnir une architecture générale de l'approche proposée pour le regroupement des articles scientifiques.

Chapitre 4 : Expérimentation et évaluation : Ce chapitre se concentre sur l'évaluation approfondie de notre méthode de regroupement de documents scientifiques, en mettant en évidence les résultats obtenus.L'objectif principal de cette évaluation est de quantifier la qualité et l'efficacité de notre méthode. Nous cherchons à confirmer nos hypothèses initiales concernant l'intérêt de combiner la similarité basée sur les réseaux de citation, l'utilisation des informations de structure (comme les titres et les résumés) sur la similarité entre les documents. En examinant les résultats, nous contribuons à une meilleure compréhension des performances de notre méthode de regroupement de documents scientifiques.

Chapitre 1

Généralités sur les documents scientifiques

1.1 Introduction

Les documents scientifiques ont une longue histoire remontant à l'Antiquité, avec des exemples tels que les oeuvres d'Aristote sur la biologie et la philosophie naturelle. Cependant, le développement et la diffusion des documents scientifiques modernes ont commencé à prendre forme à partir du 17ème siècle, lorsque la méthode scientifique est devenue plus largement utilisée. C'est également à cette époque que des journaux scientifiques ont commencé à être créés, tels que le "Journal des sçavans" en France et le "Philosophical Transactions of the Royal Society" en Angleterre.

Au cours des siècles suivants, la publication et la diffusion de documents scientifiques se sont intensifiées et sont devenues de plus en plus importantes pour la communication et la diffusion des connaissances scientifiques et technologiques. Avec l'essor d'Internet, la publication de documents scientifiques en ligne est devenue courante et a considérablement accéléré la diffusion de l'information scientifique dans le monde entier. Aujourd'hui, les documents scientifiques sont produits et publiés à un rythme effréné, reflétant l'importance de la recherche et de l'innovation dans le monde moderne.

Les documents scientifiques ont évolué en même temps que les pratiques scientifiques. Au fil des siècles, les scientifiques ont perfectionné leurs méthodes de recherche et ont utilisé différents outils pour collecter, analyser et interpréter les données. Les documents scientifiques ont également évolué pour refléter ces avancées, avec des techniques d'écriture et de présentation qui se sont améliorées pour communiquer de manière plus précise et efficace les résultats de la recherche.

Dans le monde moderne, les documents scientifiques sont devenus une partie essentielle de la communication scientifique, permettant aux scientifiques de partager leurs résultats avec leurs pairs et avec le public. Les publications scientifiques, telles que les revues scientifiques, les actes de conférences et les livres spécialisés, sont des canaux importants pour la diffusion des connaissances scientifiques. Les chercheurs et les étudiants utilisent également des bases de données en ligne pour accéder à des articles scientifiques et des données de recherche.

Avec l'augmentation constante du nombre de documents scientifiques produits chaque année, il est important pour les chercheurs de suivre les dernières avancées dans leur domaine en utilisant ces ressources et en contribuant à la production de nouveaux documents scientifiques. Les documents scientifiques sont ainsi au coeur de la recherche et de l'innovation scientifiques, contribuant à l'avancement des connaissances et à l'amélioration de la qualité de vie dans le monde.

1.2 Le document scientifique

1.2.1 Qu'est-ce qu'un document scientifique ?

Les documents scientifiques sont des documents écrits qui présentent des recherches scientifiques, des découvertes ou des théories. Ces documents peuvent prendre de nombreuses formes différentes, notamment des documents de recherche, des articles de revues, des actes de conférence, des rapports techniques et des mémoires, entre autres. Les documents scientifiques contiennent généralement une description claire et concise des méthodes de recherche, des résultats et des conclusions, souvent avec des données détaillées, des tableaux et des graphiques pour étayer les résultats. Ils sont généralement rédigés dans un style formel et technique, en utilisant une terminologie et un jargon scientifiques précis.

Les documents scientifiques sont généralement écrits pour communiquer de nouvelles découvertes, avancées ou connaissances à d'autres chercheurs, universitaires et experts dans le domaine. Ils sont souvent évalués par des pairs, ce qui signifie qu'ils sont évalués par un groupe d'experts dans le domaine pour leur exactitude, leur fiabilité et leur importance.

Les documents scientifiques jouent un rôle essentiel dans l'avancement des connaissances scientifiques et l'orientation des recherches futures. Ils sont également essentiels

pour la communication et la collaboration scientifiques, permettant aux chercheurs de partager leurs découvertes et de s'appuyer sur le travail d'autres personnes dans le domaine.[11][12]

1.2.2 Caractéristiques de la rédaction scientifique

L'écriture scientifique est une forme d'écriture unique qui se caractérise par plusieurs caractéristiques distinctes qui la distinguent des autres types d'écriture. Voici quelques-unes des principales caractéristiques de la rédaction scientifique :

1.2.2.1 Précision et exactitude

La rédaction scientifique exige précision et exactitude, car les chercheurs doivent transmettre leurs conclusions de manière claire et concise, sans laisser de place à l'ambiguïté. Cela nécessite une attention particulière aux détails, une analyse rigoureuse des données et l'utilisation d'une terminologie technique appropriée.[13][14]

1.2.2.2 Objectivité

La rédaction scientifique doit être objective et impartiale. Les chercheurs doivent éviter les opinions personnelles ou les préjugés et présenter leurs conclusions et interprétations sur la base des preuves et des données qu'ils ont recueillies.[13][14]

1.2.2.3 Clarté et concision

La rédaction scientifique doit être claire et concise. Le langage doit être simple et direct, et le texte doit être organisé de manière logique et structurée. Cela permet aux lecteurs de comprendre rapidement et facilement la recherche et sa signification.[13][14]

1.2.2.4 Utilisation des données et des preuves

La rédaction scientifique s'appuie fortement sur les données et les preuves pour étayer les affirmations et les conclusions. Les chercheurs doivent utiliser des méthodes statistiques appropriées pour analyser les données et les présenter sous forme de tableaux, de figures ou de graphiques.[13][14]

1.2.2.5 Reproductibilité

La rédaction scientifique met l'accent sur la reproductibilité, ce qui signifie que les chercheurs doivent fournir des informations détaillées sur les méthodes, l'équipement et les matériaux utilisés dans l'étude. Cela permet à d'autres chercheurs de reproduire l'étude et de vérifier les résultats.[13][14]

1.2.2.6 Citation et référence

La rédaction scientifique nécessite l'utilisation de citations et de références pour reconnaître le travail d'autres chercheurs et éviter le plagiat. Les chercheurs doivent citer leurs sources de manière précise et cohérente,en utilisant un style de citation spécifique, tel que APA ou MLA.[13][14]

1.2.2.7 Processus d'examen rigoureux

La rédaction scientifique est soumise à un processus d'examen rigoureux, avec des pairs examinateurs évaluant la recherche pour l'exactitude, la validité et la qualité. Cela garantit que la recherche répond à des normes élevées et contribue à la communauté scientifique.[13][14]

1.2.3 Types de documents scientifiques

Il existe de nombreux types de documents scientifiques, et le type spécifique de document dépendra de l'objectif et du public de la recherche. Voici quelques-uns des types de documents scientifiques les plus courants :

1.2.3.1 Articles de recherche

Il s'agit du type de document scientifique le plus courant et sont publiés dans des revues universitaires. Ils rendent compte de la recherche originale et suivent une structure spécifique, comprenant un résumé, une introduction, des méthodes, des résultats, une discussion et des références.[15]

1.2.3.2 Articles de revue

Ces documents résument et analysent l'état actuel de la recherche sur un sujet particulier. Ils fournissent généralement un large aperçu du domaine et mettent en évidence les principales conclusions, tendances et lacunes dans les connaissances.[15]

1.2.3.3 Lettres

Ce sont de courts documents qui présentent des découvertes nouvelles, intéressantes ou controversées. Ils sont généralement publiés dans un format plus court et sont souvent plus axés sur une constatation ou une observation spécifique.[15]

1.2.3.4 Actes de conférence

Ces documents sont publiés après une conférence scientifique et contiennent des résumés, des résumés ou des articles complets des recherches présentées lors de la conférence.[15]

1.2.3.5 Rapports techniques

Ce sont des documents détaillés qui décrivent les résultats de la recherche ou fournissent des informations sur des questions techniques. Ils sont généralement préparés pour un public spécifique, comme les décideurs politiques ou les professionnels de l'industrie.[15]

1.2.3.6 Thèses et mémoires

Ce sont des documents longs qui sont généralement préparés dans le cadre d'un programme d'études supérieures. Ils rendent compte de la recherche originale menée par l'étudiant et sont souvent plus détaillés et complets que les articles de recherche .[15]

1.2.3.7 Brevets

Ces documents décrivent une invention ou un procédé et sont utilisés pour protéger la propriété intellectuelle. Ils fournissent une description détaillée de l'invention et de son fonctionnement. [15]

Ce ne sont là que quelques exemples des nombreux types de documents scientifiques. Chaque type de document a un objectif et un public différent et est rédigé dans un style et un format spécifique.

1.2.4 Objectifs et finalités des documents scientifiques

Le but des documents scientifiques est de partager les connaissances scientifiques avec d'autres chercheurs et praticiens du domaine. Ils permettent aux chercheurs de communiquer leurs découvertes, théories et idées à un public plus large, favorisant ainsi la collaboration, le progrès scientifique et l'innovation.

Voici quelques finalités spécifiques des documents scientifiques :

1.2.4.1 Rapporter les résultats de la recherche

L'un des principaux objectifs des documents scientifiques est de rapporter les résultats de la recherche originale. Les chercheurs utilisent des documents scientifiques pour présenter les méthodes, les résultats et les conclusions de leurs études, et pour partager leurs découvertes avec d'autres dans le domaine. [15]

1.2.4.2 Faire progresser les connaissances scientifiques

Les documents scientifiques sont un outil essentiel pour faire progresser les connaissances scientifiques. En publiant les résultats de leurs recherches, les chercheurs peuvent contribuer à l'ensemble des connaissances existantes et aider à identifier de nouvelles orientations de recherche. [15]

1.2.4.3 Partage d'idées et de théories

Les documents scientifiques servent également de moyen de partage d'idées et de théories avec d'autres chercheurs. Grâce à des publications, les chercheurs peuvent proposer de nouvelles théories, méthodes ou approches de recherche et recevoir des commentaires d'autres experts dans le domaine. [15]

1.2.4.4 Promouvoir la collaboration

Les documents scientifiques peuvent également faciliter la collaboration entre chercheurs, car ils peuvent aider à identifier des collaborateurs ou des partenaires de recherche potentiels. Les chercheurs peuvent également utiliser des documents scientifiques pour s'appuyer sur les travaux d'autres personnes dans le domaine et collaborer à des projets de recherche conjoints.[15]

1.2.4.5 Établir la crédibilité

Les documents scientifiques sont un moyen d'établir la crédibilité et la réputation dans le domaine. La publication de recherches de haute qualité dans des revues réputées peut améliorer le profil et la réputation d'un chercheur dans le domaine, et peut conduire à des opportunités de recherche, de financement ou d'avancement de carrière supplémentaires.

Pour cela, les documents scientifiques jouent un rôle essentiel dans la communauté scientifique, servant de moyen de partage des connaissances, de faire avancer la recherche et de promouvoir la collaboration[15].

1.2.5 Structure des documents scientifiques

1.2.5.1 Le résumé

La plupart des articles sont accompagnés d'un résumé, habituellement situé en première page. Ce résumé permet aux lecteurs de saisir rapidement l'objectif de l'article et le sujet qui y est traité. En anglais, il est traduit par "abstract" ou "summary" afin de le rendre accessible à la communauté internationale. Il s'agit d'une représentation concise et fidèle du contenu de l'article, visant à informer et à susciter l'intérêt des lecteurs. Il est primordial de ne pas négliger cette partie et d'éviter toute exagération quant à l'importance des informations présentées. Généralement, les résumés sont composés d'environ 5 à 20 lignes.[16]

1.2.5.2 Les mots-clés

En complément du résumé, il est courant d'ajouter une liste de mots-clés. Ces mots-clés jouent un rôle essentiel dans le référencement et la recherche en ligne de l'article. Les revues scientifiques exigent généralement entre 3 et 10 mots-clés par article. Ces mots-clés permettent de décrire le contenu de l'article en utilisant une combinaison de termes spécifiques et de termes plus généraux liés au domaine d'étude. Ils contribuent ainsi à rendre l'article plus facilement repérable et accessible. Il est important de formuler ces mots-clés de manière originale afin d'éviter le plagiat et de refléter de manière précise le contenu de l'article.[16]

1.2.5.3 Introduction

L'introduction joue un rôle essentiel en fournissant au lecteur les fondements du sujet et en expliquant la raison d'être de la recherche. Elle établit le cadre de l'information en définissant l'axe de recherche et conduit à la problématisation. L'introduction peut être composée de plusieurs parties distinctes :[16]

• Le cadre théorique ou contexte : Cette partie consiste à décrire les connaissances actuelles sur le sujet, les découvertes récentes et les principales références en lien avec l'article.

• La problématisation : Cela implique de poser une question de recherche pertinente qui nécessite une exploration approfondie. La problématisation doit avoir des objectifs de recherche précis et peut engendrer des réactions, tels que des débats ou des questionnements.

• La formulation d'une ou plusieurs hypothèses : Il s'agit de propositions ou de pistes de réponses à vérifier ou à invalider par le biais d'expérimentations, de calculs, de sondages, d'expériences, de témoignages, de tests, etc.

1.2.5.4 Méthodologie

La section de méthodologie répond à la question "comment" dans le cadre d'une recherche scientifique. Elle constitue le cœur central de l'article, fournissant une explication détaillée des éléments clés de la recherche, des étapes de sa réalisation et de l'approche expérimentale utilisée pour vérifier les hypothèses.

Étant donné que cette partie est souvent plus développée, elle est fréquemment subdivisée en plusieurs sections distinctes. Parmi celles-ci, on retrouve notamment :[16]

• Cadre théorique (éventuellement introduit dans la section d'introduction) : Situe la recherche dans le contexte des études antérieures, en fournissant un aperçu des connaissances existantes et en établissant des liens avec d'autres travaux de recherche pertinents.

• Notation : Pour les articles impliquant l'utilisation de symboles mathématiques, cette partie vise à définir de manière originale et claire chaque symbole utilisé, afin de faciliter la compréhension des équations et des formules présentées.

• Description de la méthode : Explique en détail les différentes étapes de la méthodologie employée, permettant ainsi aux lecteurs de comprendre les expérimentations menées. Cette section précise de manière originale le rôle de chaque élément, les processus de

calcul ainsi que les principes théoriques sous-jacents à la méthode.

• Protocole expérimental : Relie de manière originale les expérimentations à l'hypothèse de départ et fournit les détails nécessaires pour reproduire l'expérience. Cela peut inclure une description originale de l'environnement expérimental, des données utilisées et d'autres informations pertinentes. Dans certains cas, le protocole expérimental peut être présenté dans la même section que la présentation des résultats, de manière originale et détaillée.

1.2.5.5 Résultats

Dans cette section, les résultats de la recherche sont présentés. Ils peuvent être représentés sous forme de tableaux, de schémas ou de graphiques pour faciliter leur analyse. La partie des résultats peut être subdivisée en sous-sections logiques (par exemple, une section par expérience menée) afin de répondre de manière optimale à la question de recherche.

Tous les résultats, y compris ceux qui ne confirment pas l'hypothèse ou qui sont neutres, doivent être inclus. La mention des résultats non concluants démontre la transparence et l'honnêteté de l'auteur envers les lecteurs et les évaluateurs.[16]

1.2.5.6 Discussion

Après la présentation des résultats, une section de discussion permet d'analyser et d'interpréter ces résultats d'un point de vue scientifique. Dans certaines revues, cette discussion peut être intégrée à la section des résultats.

La discussion permet à la recherche de se situer par rapport à d'autres études menées sur le même sujet. Elle s'appuie largement sur la littérature existante et les références bibliographiques.

De plus, la discussion éclaire les points spécifiques tels que les limites des résultats ou les domaines nécessitant des investigations plus approfondies, en anticipant les questions des lecteurs et du comité éditorial. Cependant, aucune nouvelle information n'est introduite, et toutes les étapes précédentes doivent être déjà connues du lecteur. La discussion offre une nouvelle perspective concise et précise sur les résultats, en proposant une interprétation et une analyse approfondies.[16]

1.2.5.7 Conclusion

La conclusion revêt une importance primordiale en dressant un bilan exhaustif, en résumant les résultats et les interprétations clés de la recherche. Elle offre également une opportunité de répondre à la question de recherche à la lumière des résultats obtenus, ce qui contribue à l'avancement des connaissances.

Par ailleurs, la conclusion permet d'envisager les perspectives de recherche future sans introduire de nouvelles informations. Ces travaux à venir visent à pallier les limites de l'étude présentée dans l'article et s'inscrivent dans une démarche progressive où chaque article représente une étape spécifique. Ainsi, la conclusion ouvre la voie à de nouvelles investigations et encourage une démarche de recherche progressive et continue.[16]

1.2.5.8 Remerciements

Les remerciements sont une partie de l'article qui n'est pas obligatoire. Elle est généralement placée entre la conclusion et la bibliographie, et elle permet à l'auteur de témoigner sa reconnaissance envers les personnes ou les institutions qui ont contribué à la réalisation de l'article.[16]

1.2.5.9 Bibliographie

La bibliographie contient une compilation des sources utilisées dans l'article, telles que des articles, des thèses et d'autres publications. Elle inclut toutes les références citées dans l'article qui sont directement liées au sujet, sans s'en éloigner. La bibliographie suit généralement le style, l'organisation et le format spécifiques (comme APA, Vancouver) requis par la revue ou les directives de citation.[16]

1.2.5.10 Annexes

Il est possible d'ajouter des annexes à un article scientifique, qui consistent en des tableaux, des images, des figures ou des schémas accompagnés de légendes ou de courtes descriptions. Contrairement aux éléments intégrés dans le corps principal de l'article, ceux placés en annexe ne sont pas indispensables à la compréhension de l'article, mais fournissent des informations supplémentaires et soutiennent la conclusion.[16]

La figure suivante illustre la structure d'un document scientifique :1.1

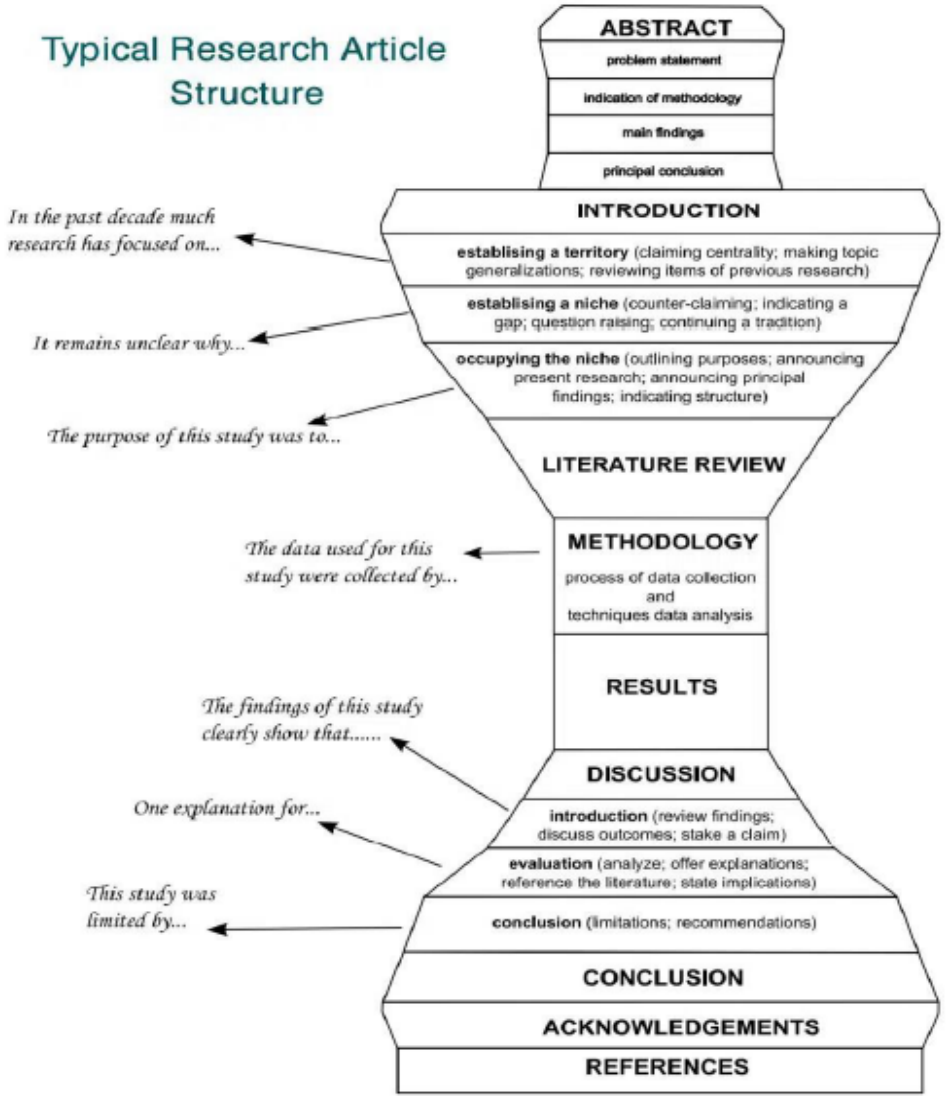

FIGURE 1.1 – La structure d'un document scientifique[1]

1.3 Les citations dans les documents scientifiques

1.3.1 Qu'est-ce que la citation

Une citation est une référence à une source d'information, telle qu'un livre, un article ou un site Web, qui a été utilisée dans la création d'un travail écrit. Les citations sont généralement incluses dans des articles universitaires, des essais et d'autres travaux pour donner crédit aux auteurs originaux et permettre aux lecteurs de localiser les sources originales.

Les citations incluent généralement des informations telles que le nom de l'auteur, le titre de l'oeuvre, la date de publication et les numéros de page où l'information a été

trouvée. Le format de citation spécifique peut varier en fonction de la discipline universitaire ou du guide de style utilisé.

Les citations ont plusieurs objectifs importants, notamment donner crédit aux auteurs originaux, établir la crédibilité des arguments de l'auteur en démontrant les sources d'information utilisées et permettre aux lecteurs de suivre les informations et les idées présentées dans l'ouvrage.[17]

1.3.2 Objectifs des citations

Les citations sont une partie importante des documents scientifiques car elles permettent à l'auteur de fournir des preuves et de légitimer les affirmations qu'il avance dans son travail. En citant les travaux précédents d'autres auteurs, l'auteur montre qu'il a étudié et compris le travail antérieur sur le sujet et qu'il peut situer son propre travail dans le contexte plus large de la recherche.

Les citations dans les documents scientifiques ont également un certain nombre d'autres fonctions importantes, notamment :

1.3.2.1 Crédibilité

Les citations aident à établir la crédibilité de l'auteur et de son travail. Les travaux qui sont largement cités dans la littérature scientifique sont considérés comme importants et influents, et les auteurs qui citent ces travaux peuvent également bénéficier de leur crédibilité.[15]

1.3.2.2 Éthique

En citant les travaux des autres, les auteurs respectent les normes éthiques de la recherche scientifique. La citation appropriée de sources est considérée comme une pratique éthique fondamentale, car elle reconnaît les contributions des autres chercheurs et évite le plagiat. [15]

1.3.2.3 Identification des sources

Les citations permettent aux lecteurs de suivre les sources mentionnées dans un document scientifique, de vérifier les informations et de trouver des sources supplémentaires pour leur propre travail de recherche. [15]

1.3.2.4 Développement de la recherche

Les citations aident à développer la recherche en fournissant un cadre pour la continuité de la recherche scientifique. Les chercheurs peuvent ainsi construire sur le travail des autres, réfuter des affirmations précédentes, ou bien suggérer de nouvelles idées ou directions de recherche.[15]

En somme, les citations sont un élément clé de la recherche scientifique qui permettent de construire des connaissances plus précises, crédibles et robustes, et qui contribuent à la continuité de la recherche dans chaque domaine scientifique.

1.3.3 les différents formats de citation

Il existe plusieurs formats de citation couramment utilisés dans les documents scientifiques, dont voici les plus courants :

1.3.3.1 Les normes APA

Les normes APA (American Psychological Association) sont largement répandues comme style de citation pour référencer les sources. Elles sont principalement utilisées par les étudiants, les chercheurs et les professeurs. Le générateur de sources APA de Scribbr offre la possibilité de citer automatiquement et gratuitement vos sources selon les normes APA, que ce soit dans le texte ou dans la bibliographie.[18][19]

Les normes APA se composent de deux règles de citation principales :

1. Citer la source dans le texte La citation courte doit être intégrée dans la phrase où l'information est utilisée. Elle se compose uniquement du nom de l'auteur et de l'année de publication (parfois accompagnée de la page si la citation concerne un passage spécifique).

2. Citer la source dans la bibliographie outes les informations détaillées de la source sont répertoriées dans la bibliographie, qui est généralement située juste après la conclusion. La référence bibliographique fournit tous les éléments nécessaires pour identifier précisément le document.

Format de bibliographie aux normes APA

Lorsque vous utilisez les normes APA, il est important de respecter certaines règles de mise en forme pour la bibliographie.

Voici les exigences de base :

• Utilisez un double interligne dans le texte.

• Définissez des marges de 2,54 cm.

• Les références doivent avoir une indentation (alinéa) de 1,27 cm.

• Il est recommandé d'utiliser la police Times New Roman en taille 12 ou Arial en taille 11 (sous réserve de l'approbation de votre institution ; d'autres polices peuvent être autorisées).

1.3.3.2 MLA

Le style de citation MLA (Modern Language Association) est largement adopté par les universitaires, en particulier dans le domaine des sciences humaines.[18][19]

Avec le style de citation MLA, vous devez inclure la source de deux manières :

1. Dans la liste des travaux cités (bibliographie), en fournissant tous les détails nécessaires pour identifier la source.

2. Dans le texte en indiquant le nom de l'auteur et le numéro de la page.

Format de la bibliographie MLA

Les travaux cités sont placés à la fin de votre document et suivent une mise en page spécifique selon le format MLA :

• La page est intitulée "Liste des travaux cités", centrée et en texte simple (sans italique, gras ou soulignement).

• Les références sont classées par ordre alphabétique du nom de famille de l'auteur.

• Utilisez un alignement à gauche et un double interligne (sans espace supplémentaire entre les références).

• Les références qui dépassent une ligne utilisent un retrait suspendu (alinéa).

• Incluez un en-tête contenant votre nom de famille et le numéro de page dans le coin supérieur droit.

1.3.3.3 Chicago

Le style de citation Chicago (notes et bibliographie) est largement utilisé dans les domaines des arts et des sciences humaines pour la rédaction de travaux académiques. Il comprend des directives concernant le format de vos documents ainsi que les techniques de citation des sources dans le texte et la bibliographie, qui est la liste des ouvrages cités.

Lorsque vous utilisez le style de citation Chicago, la première étape est de déterminer si vous devez utiliser le style Chicago A ou le style Chicago B, en fonction des directives spécifiques à votre domaine ou à votre institution.

Le style Chicago A est couramment utilisé dans les sciences humaines. Il implique l'utilisation de notes de bas de page pour les références et d'une bibliographie pour répertorier toutes les sources citées.

Le style Chicago B est un système de citation couramment utilisé dans les sciences dures. Il consiste à mentionner l'auteur et la date de publication dans le texte pour référencer les sources.

Les deux systèmes requièrent la création d'une bibliographie ainsi que la citation des sources dans le texte ou en notes de bas de page.[18][19]

1.3.3.4 IEEE

Le style IEEE (Institute of Electrical and Electronics Engineers) est principalement utilisé dans les domaines de l'ingénierie et de la technologie. La dernière édition de ce style date de 2009. Selon le manuel de style éditorial de l'IEEE, les références sont numérotées selon l'ordre des citations dans le texte. Voici quelques caractéristiques propres à ce style :

- Le prénom de l'auteur (ou ses initiales) est premier.
- Les titres des articles, des chapitres, etc. sont entre guillemets.
- Les titres des revues et livres sont en italique.
- Les titres des revues et des conférences sont abrégés.
- La bibliographie tienne les références en ordre d'apparition dans le texte.
- Utilisez « Auteur et al. » avec 4 ou plus des auteurs.[18]

1.3.3.5 Vancouver

Le style Vancouver est un style de citation numérique utilisé principalement dans le domaine médical et biomédical. Dans ce style, les références sont numérotées dans l'ordre de leur mention dans le texte, et elles sont identifiées dans le texte, les tableaux et les légendes d'illustrations à l'aide de chiffres arabes entre parenthèses.[18]

1.3.4 Contenu d'une citation

Le contenu de citation dans les documents scientifiques fait référence aux sources qui ont été utilisées pour soutenir les arguments et les conclusions présentés dans le document. Les citations sont généralement présentées sous forme de références bibliographiques et doivent être précises et complètes afin de permettre aux lecteurs de retrouver facilement les sources originales.

Le contenu d'une citation peut varier en fonction du contexte et de la source citée. En général, une citation contient les informations suivantes :

1.Nom de l'auteur ou des auteurs : le(s) nom(s) de l'auteur ou des auteurs de la publication citée doivent être mentionnés.

2.Titre de la publication : le titre complet de la publication doit être inclus. Dans le cas d'un article de revue, le titre de l'article doit être inclus en plus du titre de la revue.

3.Source de la publication : la source de la publication doit être indiquée, généralement sous la forme d'un nom de revue, d'un titre de livre ou d'un site web.

4.Date de publication : la date de publication de la publication citée doit être mentionnée.

5.Numéro de volume ou de page : si la publication est un article de revue, le numéro de volume et/ou de page doit être inclus.

6.DOI ou URL : si la publication dispose d'un DOI (Digital Object Identifier) ou d'une URL, ces informations doivent être incluses pour permettre aux lecteurs de consulter facilement la source.

En utilisant des citations appropriées dans les documents scientifiques, les auteurs peuvent soutenir leurs arguments avec des preuves solides et permettre aux lecteurs de vérifier les sources et de développer leur propre compréhension du sujet.[15]

1.3.5 Réseaux de citation

Un réseau de citations est un type de réseau qui représente les relations entre les publications savantes à travers leurs citations. Dans un réseau de citations, chaque noeud représente une publication (comme un article de revue ou un livre) et les bords entre les noeuds représentent les citations entre ces publications. Par exemple, si la publication A cité la publication B, il y aurait un bord reliant les noeuds représentant ces publications.

Les réseaux de citations peuvent être analysés pour comprendre le flux d'idées et

d'influence entre les publications et les auteurs. Ils peuvent également être utilisés pour identifier des publications importantes ou influentes dans un domaine, ou pour suivre le développement d'idées au fil du temps.

Les réseaux de citations peuvent être visualisés à l'aide de diagrammes de réseau, qui montrent les connexions entre les noeuds et peuvent aider les chercheurs à identifier des modèles et des relations dans les données. L'analyse des réseaux de citations est devenue de plus en plus importante à l'ère numérique, car elle permet aux chercheurs d'analyser rapidement et efficacement de grandes quantités de données scientifiques.[20]

1.3.6 les types des réseaux de citation

Il existe plusieurs types de réseaux de citations qui peuvent être analysés dans la recherche universitaire. Cette figure représente quelques types courants :1.2

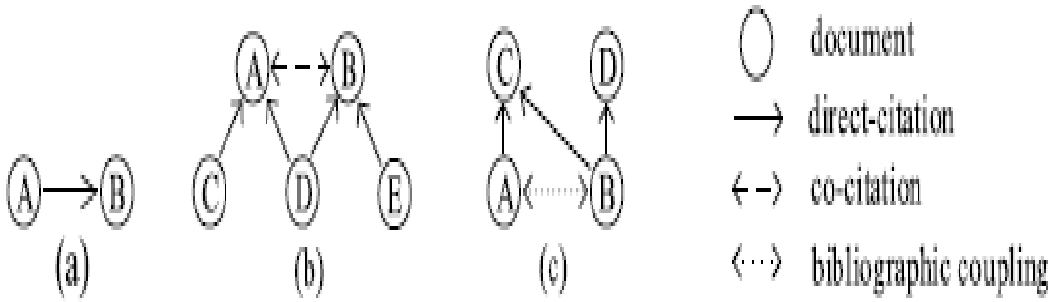

FIGURE 1.2 – Exemple d'un réseaux de citation[2]

1.3.6.1 Direct citation network

Dans un réseau de citation directe, deux publications sont liées si l'une cite directement l'autre. Ce type de réseau peut être utilisé pour analyser l'influence d'une publication ou d'un auteur particulier sur un domaine plus large.[21]

1.3.6.2 Co-citation network

Dans un réseau de Co-citation, deux publications sont liées si elles sont toutes les deux citées par une troisième publication. Cela peut aider les chercheurs à identifier des

groupes de publications connexes ou à cartographier la structure intellectuelle d'un domaine particulier.[21]

1.3.6.3 Bibliographic coupling network

Dans un réseau de couplage bibliographique, deux publications sont liées si elles citent toutes deux la même troisième publication. Ce type de réseau peut être utilisé pour identifier les publications importantes ou influentes dans un domaine.[21]

Il existe de nombreux autres types de réseaux de citations qui peuvent être analysés, et les chercheurs peuvent choisir d'utiliser différents types de réseaux en fonction de leurs questions de recherche et de leurs données.

1.4 Conclusion

Pour conclure ce premier chapitre, nous avons examiné les diverses dimensions des documents scientifiques, y compris leurs caractéristiques, leurs types et leur structure. Nous avons également abordé les citations et leurs éléments essentiels, en mentionnant les différents formats et types de citation, ainsi que les réseaux de citation qui permettent de visualiser les relations entre les documents scientifiques. Cette compréhension approfondie jettera des bases solides pour le deuxième chapitre, où nous explorerons le domaine passionnant de l'apprentissage automatique et du clustering.

Chapitre 2

Classification automatique des documents

2.1 Introduction

Après avoir présenté le document scientifique dan le premier chapitre, nous aborderons dans ce deuxième chapitre la problématique de la classification automatique de document textuels. Même si la problématique de la classification automatique du texte date de plusieurs années, elle est toujours d'actualité et trouve actuellement un énormément regain d'intérêt dans la communauté de la recherche scientifique. Les gros volumes de documents textuels générés continuellement d'une part, et l'arrivée de la communauté de l'apprentissage automatique dans ce domaine ont contribué grandement au regain d"intérêt pour cette problématique.

En effet devant des énormes quantités de documents tout le temps en croissance, on a besoin d'outils qui peuvent aider les utilisateurs à se repérer, à organiser ces collections de documents, à trouver facilement les documents qui les intéressent. La classification automatique de document repend à cette problématique. Elle consiste à classer d'une manière automatique une collection de documents selon des critères (type de document, style du texte, thème etc.) dans le cadre d'une classification supervisée ou sinon de faire un regroupement automatique d'un ensemble de documents selon leurs contenus (sujets abordés) de ces documents dans le cadre du Clustering.

Dans ce chapitre nous rappelons brièvement le domaine de l'apprentissage automatique, nous aborderons ensuite la classification automatique de documents et nous nous focaliserons beaucoup plus sur la technique de Clustering de documents textuels.

2.2 L'apprentissage automatique ou Machine Learning

2.2.1 Qu'est-ce que le Machine Learning ?

L'apprentissage automatique (Machine Learning) est une branche de l'intelligence artificielle (l'IA) qui vise à permettre aux machines d'imiter le comportement intelligent humain. Les systèmes d'intelligence artificielle sont conçus pour résoudre des tâches complexes de manière similaire à celle des êtres humains. L'objectif principal de l'IA est de créer des modèles informatiques capables de présenter des comportements intelligents tels que la reconnaissance visuelle, la compréhension du langage naturel ou l'exécution d'actions dans le monde physique.

L'apprentissage automatique regroupe un ensemble de techniques qui sont utilisées pour réaliser cette vision de l'IA. Il a été défini dans les années 1950 par Arthur Samuel, un pionnier de l'IA, comme étant le domaine qui donne aux ordinateurs la capacité d'apprendre sans être explicitement programmés.

Le processus d'apprentissage automatique commence par la collecte de données, telles que des chiffres, des photos, du texte ou des enregistrements provenant de diverses sources. Ces données sont préparées et utilisées comme ensemble d'entraînement, sur lequel le modèle d'apprentissage automatique sera formé. Plus les données sont abondantes, meilleure sera la performance du programme.

Ensuite, les programmeurs choisissent un modèle d'apprentissage automatique approprié, fournissent les données d'entraînement et laissent le modèle s'entraîner pour découvrir des schémas ou faire des prédictions. Au fil du temps, le programmeur peut ajuster le modèle en modifiant ses paramètres pour améliorer sa précision.

Une partie des données est conservée pour être utilisée comme données d'évaluation, permettant de tester la précision du modèle sur de nouvelles données. Le résultat final est un modèle d'apprentissage automatique qui peut être utilisé pour effectuer des tâches similaires sur différents ensembles de données à l'avenir.[22]

2.2.2 Comment fonctionne le Machine Learning ?

Le Machine Learning offre une variété de modèles qui utilisent des techniques algorithmiques spécifiques. Le développement d'un modèle de Machine Learning repose sur quatre étapes principales. cette figure illustre le fonctionnement de Machine Learning : 2.1

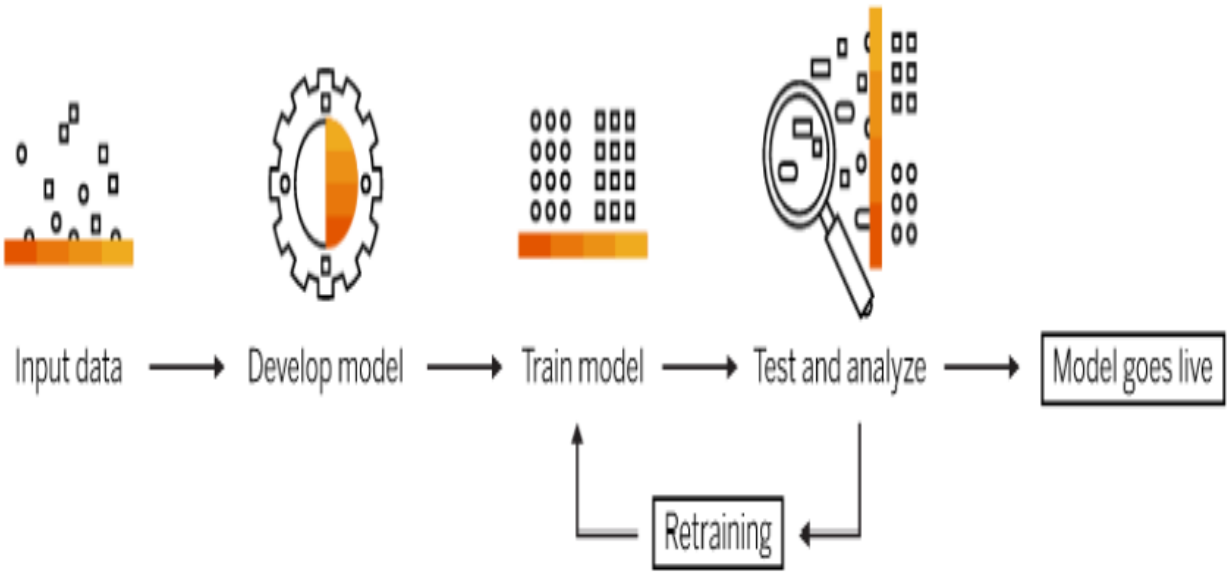

FIGURE 2.1 – Le fonctionnement de Machine Learning [3]

La première étape : consiste à sélectionner et à préparer un ensemble de données d'entraînement. Ces données seront utilisées pour nourrir le modèle de Machine Learning pour apprendre à résoudre le problème pour lequel il est conçu. Les données peuvent être étiquetées, afin d'indiquer au modèle les caractéristiques qu'il devra identifier. Elles peuvent aussi être non étiquetées, et le modèle devra repérer et extraire les caractéristiques récurrentes de lui-même. Dans les deux cas, les données doivent être soigneusement préparées, organisées et nettoyées. Dans le cas contraire, l'entraînement du modèle de Machine Learning risque d'être biaisé. Les résultats de ses futures prédictions seront directement impactés.[23]

La deuxième étape :consiste à sélectionner un algorithme à exécuter sur l'ensemble de données d'entraînement. Le type d'algorithme à utiliser dépend du type et du volume de données d'entraînement et du type de problème à résoudre.[23]

La troisième étape : La troisième étape est l'entraînement de l'algorithme. Il s'agit d'un processus itératif. Des variables sont exécutées à travers l'algorithme, et les résultats

sont comparés avec ceux qu'il aurait dû produire. Les paramètres peuvent ensuite être ajustés pour accroître la précision du résultat. On exécute ensuite de nouveau les variables jusqu'à ce que l'algorithme produise le résultat correct la plupart du temps. L'algorithme, ainsi entraîné, est le modèle de Machine Learning.[23]

La quatrième et dernière étape :est l'utilisation et l'amélioration du modèle. On utilise le modèle sur de nouvelles données, dont la provenance dépend du problème à résoudre. Par exemple, un modèle de Machine Learning conçu pour détecter les spams sera utilisé sur des emails. [23]

2.2.3 Les modèles de Machine Learning

La figure ci-dessous montre les differents modèles de Machine Learning :2.2

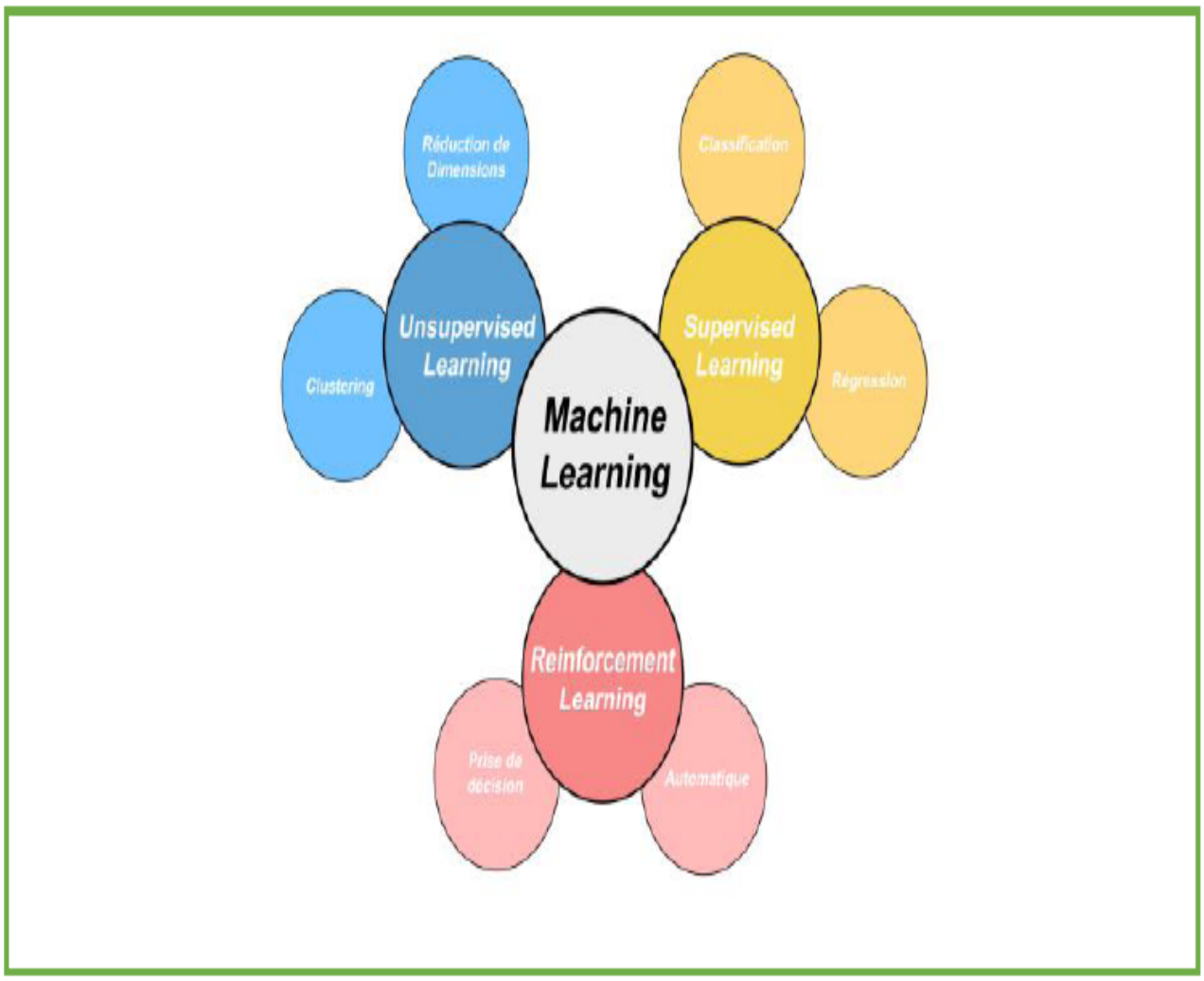

FIGURE 2.2 – Les modèles de Machine Learning [4]

Il existe plusieurs modèles ou algorithmes d'apprentissage automatique, ils sont généralement classés dans trois catégories :

- Apprentissage supervisé
- Apprentissage non supervisé
- Apprentissage par renforcement

2.2.3.1 Apprentissage automatique supervisé

Dans l'apprentissage automatique supervisé, les modèles sont entraînés à l'aide d'ensembles de données étiquetés, ce qui leur permet d'apprendre et de s'améliorer au fil du temps. Ce type d'apprentissage consiste à fournir à la machine des exemples de données accompagnés de leurs étiquettes correspondantes. Par exemple, un algorithme peut être entraîné à reconnaître des images radiologiques correspondant à un type de cancer en lui présentant un ensemble d'images étiquetées comme "cancer" et d'autres étiquetées comme « non cancer ». Grâce à ces exemples, la machine apprend à identifier les caractéristiques distinctives des images de cancer et peut ensuite généraliser pour reconnaître d'autres images de cancer par elle-même. Ainsi, au fur et à mesure de l'entraînement, le modèle devient plus précis dans sa capacité à classifier les images correctement. C'est un exemple très simple de classification automatique.[22]

2.2.3.2 Apprentissage automatique non supervisé

L'apprentissage non supervisé est une technique qui consiste à entraîner des modèles sans utiliser préalablement d'étiquetage manuel ou automatique des données. L'objectif de l'apprentissage non supervisé est de découvrir des motifs et des structures cachées dans les données. Contrairement à l'apprentissage supervisé, il n'y a pas de réponse connue à prédire dans l'apprentissage non supervisé.

Les algorithmes d'apprentissage non supervisé regroupent les données en fonction de leur similitude, sans intervention humaine pour les guider. Ces algorithmes analysent les caractéristiques des données et identifient des schémas, des clusters ou des associations significatives entre les différentes instances de données. Cette approche permet de découvrir des informations précieuses et des insights sur les données, sans avoir besoin d'une étiquette préalable pour chaque donnée.

L'apprentissage non supervisé diffère de l'apprentissage supervisé en termes de données

d'entrée. Dans l'apprentissage supervisé, le modèle reçoit des données d'entraînement étiquetées, où les entrées sont associées à des sorties attendues. En revanche, dans l'apprentissage non supervisé, le modèle apprend à partir de données d'entraînement non étiquetées afin de découvrir des motifs et des structures.

En apprentissage non supervisé, le clustering est l'algorithme le plus utilisé. Il consiste à regrouper des données hétérogènes en groupes de données ayant des caractéristiques homogènes.[24]

2.2.3.3 Apprentissage automatique par renforcement

L'apprentissage automatique par renforcement consiste à entraîner des machines à prendre des actions optimales en utilisant un système de récompense basé sur des essais et des erreurs. Ce type d'apprentissage est souvent utilisé pour former des modèles à jouer à des jeux ou à entraîner des véhicules autonomes à conduire de manière efficace.

Dans l'apprentissage par renforcement, la machine interagit avec son environnement et effectue des actions. En fonction de ces actions, elle reçoit des récompenses qui évaluent la qualité de ses décisions. L'objectif est d'apprendre à maximiser les récompenses obtenues au fil du temps.[22]

2.2.4 Exemples d'application de l'apprentissage automatique

L'apprentissage automatique est actuellement utilisé pour répondre à des problèmes complexes et variés dans presque tous les domaines : santé, économie, militaires, sécurité, recherche ....etc. On peut citer dans ce qui suit quelques exemples :

Détection des spams Un spam désigne toute communication non sollicitée envoyée en masse. Généralement envoyé par e-mail, le spam est également distribué par le biais de SMS, des réseaux sociaux ou d'appels téléphoniques. Les messages de spam se présentent souvent sous la forme d'e-mails promotionnels inoffensifs (bien que fastidieux). Mais parfois, le spam est une arnaque frauduleuse ou malveillante. La détection des spams est importante pour protéger les utilisateurs contre les tentatives d'escroquerie, de phishing et de malwares. Il existe plusieurs méthodes pour détecter les spams, notamment L'apprentissage automatique : utilise des algorithmes d'apprentissage pour classifier les courriers électroniques en tant que spam ou non spam, cette méthode utilise des ensembles de données de courriers électroniques connus, classés comme étant des spams ou des cour-

riers électroniques légitimes, pour entraîner un modèle d'apprentissage. Le modèle utilise des caractéristiques telles que les mots clés, la longueur du texte, la présence de liens, la présence de pièces jointes, etc. pour classer les courriers électroniques.[25]

Détection d'intrusion Un système de détection d'intrusions (IDS, de l'anglais Intrusion Detection System) est un périphérique ou processus actif qui analyse l'activité du système et du réseau pour détecter toute entrée non autorisée et / ou toute activité malveillante. La manière dont un IDS détecte des anomalies peut beaucoup varier ; cependant, l'objectif principal de tout IDS est de prendre sur le fait les auteurs avant qu'ils ne puissent vraiment endommager vos ressources. La détection d'intrusion basée sur l'apprentissage automatique (en anglais Machine Learning based Intrusion Detection System ou ML-IDS) est une technique de sécurité informatique qui utilise l'intelligence artificielle pour identifier les activités malveillantes sur un réseau ou sur un système d'information. Elle est basée sur des algorithmes d'apprentissage automatique qui apprennent à détecter les comportements anormaux en analysant les données du réseau.[26]

Diagnostic médical Les algorithmes d'apprentissage automatique ont la capacité d'exploiter de grandes quantités de données stockées afin de découvrir des connexions essentielles dans le processus de prise de décision. Des applications militaires à la médecine, l'informatique peut être utilisée pour analyser des images. Des experts en imagerie et cliniques se sont associés pour décrire l'impact de l'apprentissage automatique sur l'interprétation des images médicales, telles que les radiographies pulmonaires. Les sous-types d'apprentissage automatique, tels que les réseaux de neurones convolutifs, "peuvent identifier des changements subtils dans les radiographies pulmonaires, et dans certains cas, les niveaux de précision pour diagnostiquer des conditions, telles que la pneumonie, sont équivalents ou supérieurs à ceux des cliniciens", notent les scientifiques. "Contrairement aux méthodes statistiques traditionnelles, où les inférences sont faites en fonction de la population étudiée, les algorithmes d'apprentissage automatique imitent les processus cognitifs humains lors de la prise de décision."[27]

Prédiction de fraudes Machine Learning peut être utilisé pour la prédiction de fraudes dans divers domaines, tels que la finance, l'assurance, et le commerce électronique. Les algorithmes de machine Learning peuvent analyser les données transactionnelles et identifier des modèles qui peuvent indiquer une activité frauduleuse. Les modèles peuvent être utilisés pour prédire les transactions futures qui sont susceptibles d'être frauduleuses

et pour aider les entreprises à prendre des mesures préventives pour éviter les pertes financières.[28]

2.3 Le Clustering

2.3.1 Qu'est-ce que le clustering ?

Le clustering est une méthode d'apprentissage automatique qui consiste à regrouper des points de données par similarité ou par distance. C'est une méthode d'apprentissage non supervisée et une technique populaire d'analyse statistique des données. L'objectif est de diviser un ensemble de données en plusieurs groupes ou "clusters" de telle manière que les éléments au sein d'un même cluster soient similaires entre eux et différents des éléments des autres clusters.

Un cluster est un groupe de données similaires qui ont été regroupées ensemble par un algorithme de clustering. Les données regroupées dans un cluster partagent des caractéristiques ou des propriétés communes qui les distinguent des autres données dans l'ensemble de données.

Le nombre de clusters formés dépend du nombre de groupes de données similaires que l'algorithme de clustering est capable de détecter dans l'ensemble de données. Le nombre optimal de clusters peut être déterminé par l'utilisateur en fonction de l'objectif de l'analyse et de la nature des données.[29]

Voici un exemple de génération des clusters :2.3

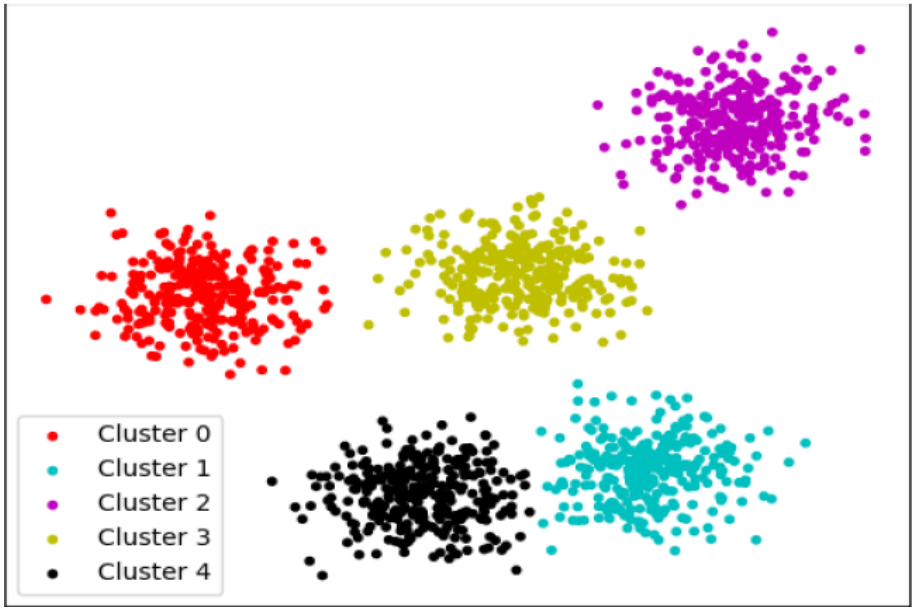

FIGURE 2.3 – Exemple de génération des clusters[5]

2.3.2 Principe de clustering

Cette méthode de classification non supervisée rassemble un ensemble d'algorithmes d'apprentissage dont le but est de regrouper entre elles des données non étiquetées présentant des propriétés similaires. Le but des algorithmes de clustering est de donner un sens aux données et d'extraire de la valeur à partir de grandes quantités de données structurées et non structurées. Il s'agit d'une technique d'analyse statistique des données très utilisée dans de nombreux domaines, y compris l'apprentissage automatique, la reconnaissance de formes, le traitement de signal et d'images, la recherche d'information, la bio-informatique, la compression de données el'infographie,etc.[29]

Son principe consiste à :

Etant donné :

- Un ensemble de points, chacun ayant un ensemble d'attributs,
- Et une mesure de similarité d définie sur cet ensemble de point,

Trouver des groupes (classes, segments, clusters) tels que :

- Les points à l'intérieur d'un même groupe sont très similaires entre eux.
- Les points appartenant à des groupes différents sont très dissimilaires.

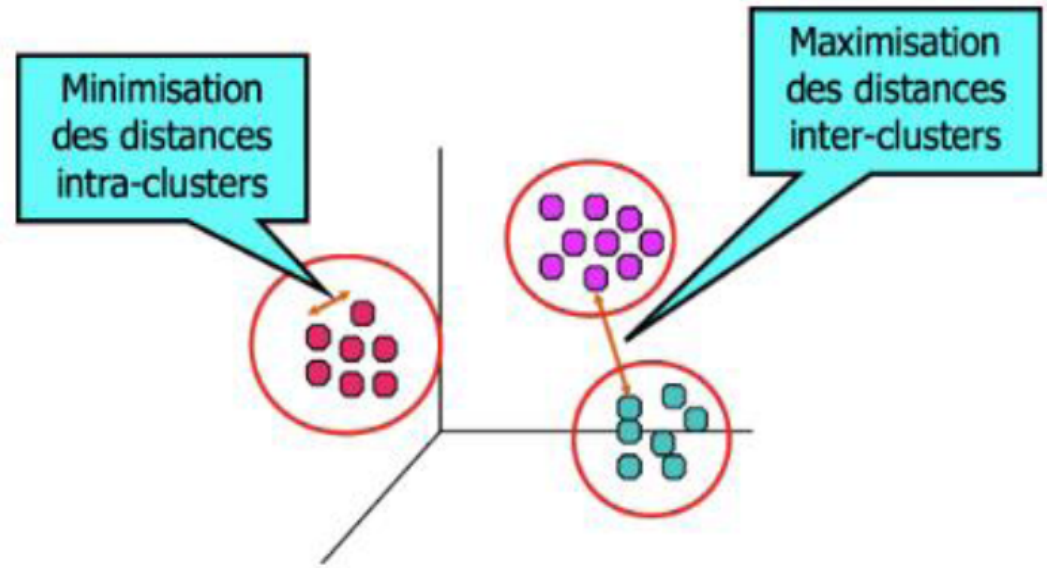

FIGURE 2.4 – Principe de clustering[6]

2.3.3 Approches de Clustering

2.3.3.1 Approche par partitionnement

Appelée aussi approche des centroïdes. La méthode centroïde la plus classique est la méthode des k-moyenne. Elle ne nécessite qu'un seul choix de départ : k, le nombre de classes voulues. On initialise l'algorithme avec k points au hasard parmi les n individus.

Ces k points représentent alors les k classes dans cette première étape. On associe ensuite chacun des n-k points restants à la « classe-point » qui lui est la plus proche. A la fin de cette première étape, chaque classe est caractérisée par la moyenne des valeurs de chacun de ses individus. On a k moyennes pour k classes. La deuxième étape consiste à évaluer la distance de chaque individu à chacune des k moyennes. Certains individus peuvent ici changer de classe.

A la fin de cette étape, on actualise les k moyennes. Et on réitère les étapes, jusqu'à ce qu'il y ait convergence pour obtenir nos k clusters finaux. Ces classes finales dépendent souvent beaucoup des k individus choisis pour l'initialisation. C'est pourquoi certains algorithmes de k-means itèrent plusieurs fois le processus avec des initialisations différentes, dans le but de garder la partition qui minimise le plus la variance intra-classes (somme des distances entre les individus d'une même classe).

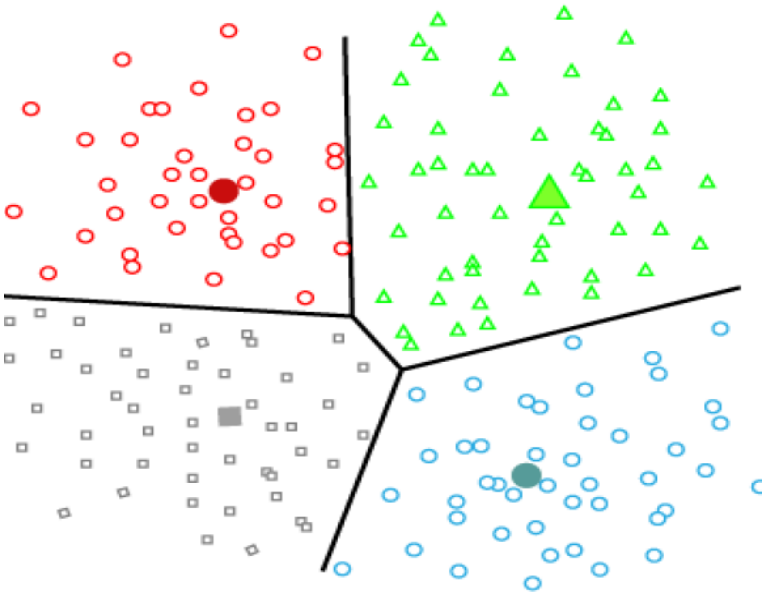

FIGURE 2.5 – Approche par partitionnement [7]

2.3.3.2 Approche hiérarchique

Dans cette méthode, une décomposition hiérarchique de l'ensemble donné d'objets de données est créée. Nous pouvons classer les méthodes hiérarchiques et nous pourrons connaître le but de la classification sur la base de la formation de la décomposition hiérarchique. Il existe deux types d'approches pour la création de décomposition hiérarchique, ce sont :

Approche agglomérative

L'approche agglomérative est également connue sous le nom d'approche ascendante. Initialement, les données sont divisées en objets qui forment des groupes séparés. Par la suite, il continue de fusionner les objets ou les groupes qui sont proches les uns des autres, ce qui signifie qu'ils présentent des propriétés similaires. Ce processus de fusion se poursuit jusqu'à ce que la condition de fin soit remplie.[30]

L'approche de division

Est également connue sous le nom d'approche descendante. Dans cette approche, nous commencerions par les objets de données qui se trouvent dans le même cluster. Le groupe de clusters individuels est divisé en petits clusters par itération continue. L'itération continue jusqu'à ce que la condition de terminaison soit remplie ou jusqu'à ce que chaque cluster contienne un objet.[30]

Voici un exemple de fonctionnement des deux approches :2.6

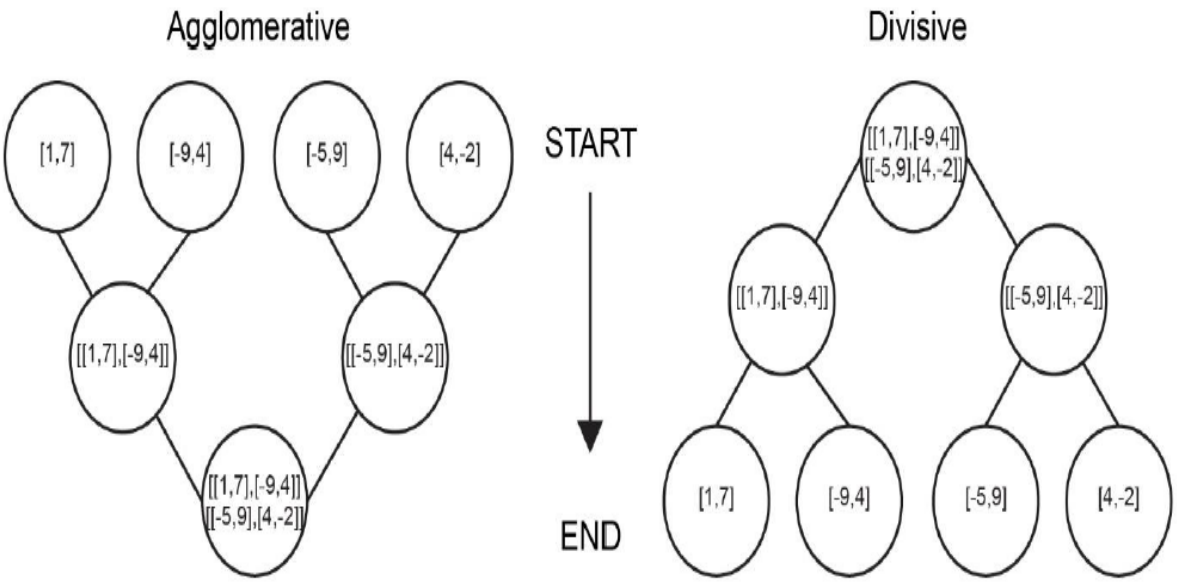

FIGURE 2.6 – Illustrastion de fonctionnement des deux approches[8]

Une fois que le groupe est divisé ou fusionné, il ne peut jamais être annulé car il s'agit d'une méthode rigide et moins flexible. Les deux approches qui peuvent être utilisées pour améliorer la qualité du clustering hiérarchique dans l'exploration de données sont :

- Il convient d'analyser soigneusement les liens de l'objet à chaque partitionnement du clustering hiérarchique.

- On peut utiliser un algorithme d'agglomération hiérarchique pour l'intégration de l'agglomération hiérarchique. Dans cette approche, dans un premier temps, les objets sont regroupés en micro-clusters. Après avoir regroupé les objets de données en micro-clusters, le macro clustering est effectué sur le micro-cluster.

2.3.3.3 Approche basée sur la densité

La méthode basée sur la densité se concentre principalement sur la densité. Dans cette méthode, le cluster donné continuera à croître de manière continue tant que la densité dans le voisinage dépasse un certain seuil, c'est-à-dire pour chaque point de données dans un cluster donné. Le rayon d'un cluster donné doit contenir au moins un nombre minimum de points.[30]

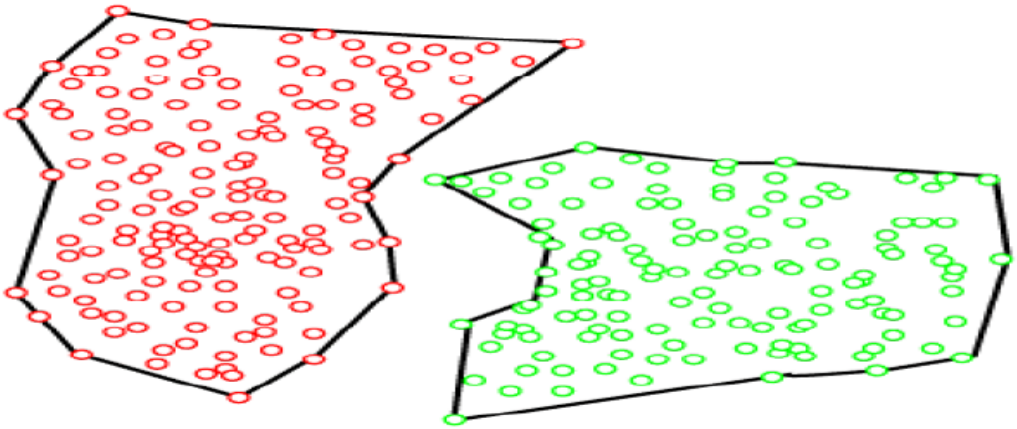

FIGURE 2.7 – Le fonctionnement de la méthode basé sur la densité[9]

2.3.4 Algorithmes de Clustering

2.3.4.1 K-Means

K-means est un algorithme de clustering largement utilisé pour regrouper un ensemble de données en K clusters distincts. L'objectif de l'algorithme est de minimiser la somme des distances entre chaque point de données et le centre de son cluster respectif. Le fonctionnement de l'algorithme est le suivant :

1. Le nombre de clusters K est choisi à l'avance par l'utilisateur.

2. Les centres de cluster initiaux sont choisis aléatoirement parmi les points de données.

3. Chaque point de données est affecté au centre de cluster le plus proche (distance Euclidienne).

4. Les centres de cluster sont recalculés en utilisant la moyenne des points de données qui y sont affectés.

5. Les étapes 3 et 4 sont répétées jusqu'à ce que les centres de cluster ne bougent plus significativement ou que le nombre maximum d'itérations soit atteint.[31]

L'algorithme converge généralement assez rapidement, mais il peut être sensible au choix initial des centres de cluster et peut aboutir à des optima locaux. Des variantes de l'algorithme, comme le K-means ++, ont été proposées pour résoudre ce problème.

Pseudo-algorithme de K-Means :

Algorithm 1 Algorithme des K-Means

Entrée : $D = \{t_1, t_2, \ldots, t_n\}$ {Ensemble d'éléments}

K {Nombre de clusters souhaités}

Sortie : K {Ensemble de clusters}

Algorithme K-Means :

Assigner des valeurs initiales pour $m_1, m_2, \ldots, m_k$

Répéter

Assigner chaque élément t_i au cluster qui a la moyenne la plus proche

Calculer la nouvelle moyenne pour chaque cluster

Jusqu'à ce que le critère de convergence soit atteint

Exemple du Fonctionnement de l'algorithme

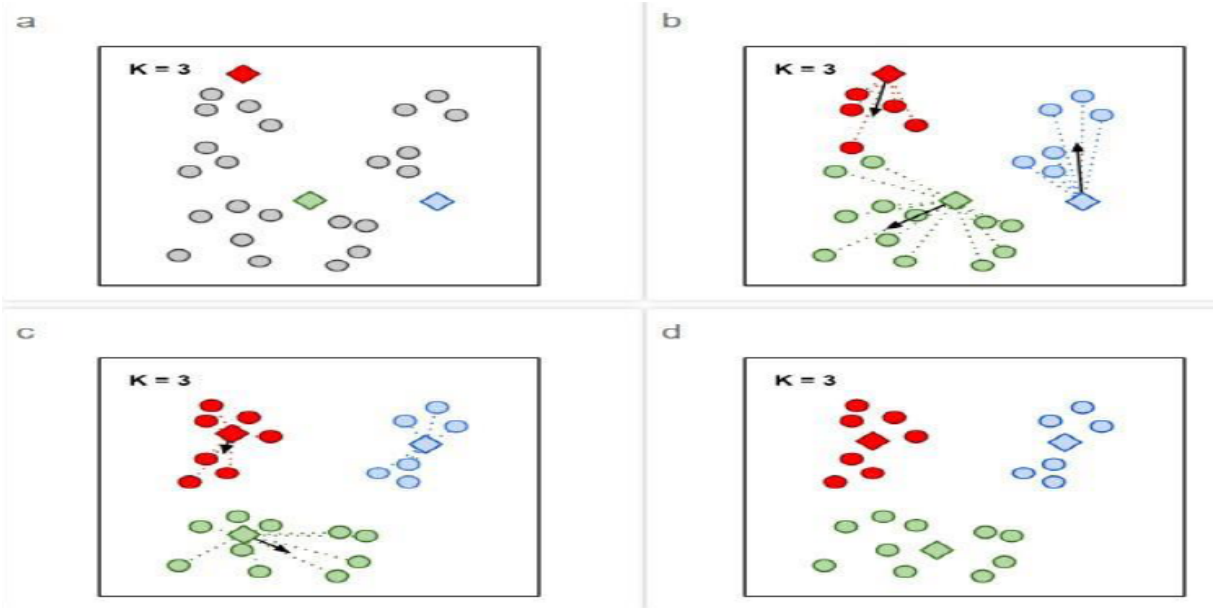

FIGURE 2.8 – Exemple du Fonctionnement de l'algorithme

Avantages de l'algorithme K-means sont les suivants

1. Rapidité : K-means est relativement rapide et peut gérer de grands ensembles de données.

2. Facilité d'utilisation : L'algorithme est simple à comprendre et à mettre en oeuvre, même pour les débutants.

3. Scalabilité : L'algorithme peut être facilement étendu pour traiter de grands ensembles de données.

4. Interprétabilité : Les clusters obtenus sont facilement interprétables, car chaque cluster est représenté par son centre de cluster.

5. Performance : K-means est généralement efficace pour séparer les données en clusters, en particulier lorsque les clusters sont bien séparés.

6. Applications diverses : L'algorithme est applicable dans de nombreux domaines, notamment la reconnaissance de formes, la segmentation d'images, la recommandation de produits, l'analyse de marché, la bioinformatique, etc.

Inconvénients de l'algorithme K-means sont les suivants

Dépendance aux paramètres : Le choix du nombre de clusters K doit être effectué à l'avance, et le résultat final peut varier en fonction de ce choix initial.

1. Sensibilité aux valeurs aberrantes : Les valeurs aberrantes peuvent considérablement affecter le résultat final, en particulier lorsqu'elles sont incluses dans un cluster.

2. Sensibilité aux centres de cluster initiaux : Les centres de cluster initiaux sont choisis aléatoirement, ce qui peut conduire à des résultats différents selon les exécutions.

3. Dépendance à l'échelle : L'algorithme est sensible à l'échelle des données, il est donc important de normaliser les données avant de les utiliser.

4. Clusters non sphériques : K-means est plus efficace pour séparer les données en clusters sphériques, il peut donc ne pas être adapté aux clusters de formes arbitraires.

5. Risque d'optimum local : L'algorithme peut converger vers un optimum local au lieu de l'optimum global, ce qui peut conduire à des résultats sous-optimaux.

2.3.4.2 DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) est un algorithme de clustering qui se base sur la densité des points pour regrouper les données en clusters. L'algorithme DBSCAN peut identifier les clusters de formes arbitraires, et est capable de traiter les données avec des valeurs aberrantes.

Le principe de base de l'algorithme DBSCAN est de regrouper les points qui sont suffisamment proches les uns des autres pour former un cluster, tout en excluant les points qui sont isolés et considérés comme du bruit. Pour ce faire, l'algorithme DBSCAN utilise deux paramètres principaux :

1. eps : la distance maximale entre deux points pour qu'ils soient considérés comme faisant partie du même cluster.

2. minPts : le nombre minimal de points nécessaires pour qu'un cluster soit considéré comme valide.

L'algorithme DBSCAN commence par choisir un point de départ et explore ensuite tous les points voisins qui se trouvent à une distance inférieure à eps. Si le nombre de voisins est supérieur ou égal à minPts, alors un nouveau cluster est formé. Si le nombre de voisins est inférieur à minPts, alors le point est considéré comme du bruit. Si le nombre de voisins est suffisant mais les voisins ne sont pas tous connectés, alors plusieurs clusters peuvent être formés.[32]

Pseudo-Algorithme de DBSCAN

Partie 1 : Algorithme DBSCAN (étape principale)

Algorithm 2 DBSCAN (Partie 1)

Input : D, ε, $MinPts$
$C \leftarrow 0$
for each unvisited point P in dataset D **do**
  mark P as visited
  $NeighborPts \leftarrow$ regionQuery(P, ε)
  if sizeof$(NeighborPts) <$ MinPts **then**
    mark P as NOISE
  else
    $C \leftarrow$ next cluster
    expandCluster(P, $NeighborPts$, C, ε, $MinPts$)
  end if
end for

Partie 2 : Fonctions expandCluster and regionQuery

Algorithm 3 DBSCAN (Partie 2)

expandCluster(P, $NeighborPts$, C, ε, $MinPts$) :
add P to cluster C
for each point P' in $NeighborPts$ **do**
  if P' is not visited **then**
    mark P' as visited
    $NeighborPts' \leftarrow$ regionQuery(P', ε)
    if sizeof($NeighborPts'$) $\geq$ MinPts **then**
      $NeighborPts \leftarrow NeighborPts \cup NeighborPts'$
    end if
  end if
  if P' is not yet a member of any cluster **then**
    add P' to cluster C
  end if
end for
regionQuery(P, ε) :
return all points within P's ε-neighborhood (including P)

Les avantages de l'algorithme DBSCAN sont les suivants

1. Robustesse aux valeurs aberrantes : L'algorithme DBSCAN est capable de traiter les données avec des valeurs aberrantes et de les exclure de manière automatique.

2. Identification de clusters de formes arbitraires : L'algorithme DBSCAN peut identifier des clusters de formes arbitraires, contrairement à l'algorithme K-means qui ne peut identifier que des clusters sphériques.

3. Pas besoin de spécifier le nombre de clusters : L'algorithme DBSCAN peut identifier le nombre de clusters de manière automatique. Cependant, l'algorithme DBSCAN peut être sensible au choix des paramètres eps et minPts, et peut être lent pour les ensembles de données très grands et/ou de haute dimension. De plus, le résultat final peut dépendre de l'ordre dans lequel les points sont explorés, ce qui peut affecter la stabilité de l'algorithme.

2.3.4.3 Mean-shift

Mean-shift est un algorithme de clustering non paramétrique qui est souvent utilisé pour identifier des clusters dans des données de haute dimension ou avec des formes complexes. L'algorithme Mean-shift utilise une technique de détection de mode qui est basée sur la densité des données.

L'algorithme Mean-shift commence par choisir un point de départ dans l'espace de données et calcule la densité locale des données dans une région autour de ce point en utilisant une fonction de noyau. La fonction de noyau sert à pondérer les contributions des points voisins à la densité locale. Ensuite, l'algorithme déplace le point de départ dans la direction de la plus grande augmentation de densité, c'est-à-dire vers le mode local de la densité. Ce processus est répété pour chaque point de départ jusqu'à ce que chaque point ait convergé vers un mode local de densité, qui représente un cluster.[33]

Pseudo-Algorithme de Mean shift :

Algorithm 4 Mean Shift

$i \leftarrow 1, y_i \leftarrow x_i$

Center a window on y_i // initialisation

repeat

$U_h(y_i)$ // assess the sample mean of data falling within the window (neighborhood of y_i depending on Euclidean distance)

$y_{i+1} \leftarrow U_h(y_i)$

Move the window from x_i to y_{i+1}

$i \leftarrow i + 1$

until Stabilization // a mode has been found

Les avantages de l'algorithme Mean-shift sont les suivants

1. Pas besoin de spécifier le nombre de clusters : L'algorithme Mean-shift est un algorithme non paramétrique qui n'exige pas que le nombre de clusters soit spécifié à l'avance.

2. Capable de traiter des formes complexes : L'algorithme Mean-shift peut identifier des clusters de formes arbitraires et est capable de traiter des données de haute dimension.

3. Pas sensible aux paramètres de distance : L'algorithme Mean-shift utilise une fonction de noyau pour pondérer les contributions des points voisins, ce qui permet d'obtenir des résultats robustes par rapport aux paramètres de distance.

Cependant, l'algorithme Mean-shift peut être lent pour les ensembles de données très grands et/ou de haute dimension, car il doit calculer la densité locale de chaque point de données à chaque itération. De plus, l'algorithme peut être sensible au choix de la bande passante de la fonction de noyau, qui peut affecter la précision des résultats.

2.3.5 Mesure de qualités du clustering

Les mesures de qualité du clustering sont utilisées pour évaluer la qualité des résultats d'un algorithme de clustering en mesurant la similarité des membres d'un cluster et la dissimilarité entre les différents clusters, On distinguera deux familles de mesures de qualité de clustering : les indices opérant sur des graphes et ceux basés sur les distances.[34][35]

2.3.5.1 Mesures de qualité opérant sur les graphes

Les indices de qualité opérant sur des graphes s'intéressent aux liens entre les noeuds à l'intérieur des graphes. Ils se basent sur le principe que les noeuds appartenant à une même classe sont plus liés entre eux qu'avec les points appartenant à des classes différentes. Plusieurs formalisations ont été développées : la « Performance » et la « Modularité »

La modularité :

La modularité est une mesure qui compare la distribution des arêtes dans un graphe par rapport à la probabilité qu'il y ait un arc entre deux noeuds. Autrement dit, la modularité compare la proximité des noeuds à celle attendue par hasard (Newman 2006). Cette mesure combine une comparaison entre la partition et l'hypothèse nulle pour déterminer si la partition est significative. [34]

Voici la formulation de Newman pour le calcul de la modularité.

$$Q = \frac{1}{2m} \sum_{i=1}^{i=n} \sum_{j=1}^{j=n} (A_{ij} - \frac{K^i * K_j}{2m}) \wp(C_i, C_j) \tag{2.1}$$

La Performance :

La performance introduit la notion de couple de noeuds correctement interprétés qui désigne à la fois les couples de noeuds liés appartenant à une même classe et les couples de noeuds non liés appartenant à deux classes différentes. La Performance calcule la fraction des couples de noeuds correctement interprétés par rapport au nombre total de couples de noeuds :[34]

$$\text{Performance}(p) = \frac{m(p) + \sum_{u,v \notin E, u \in C_i, v \in C_j, i \neq j}}{\frac{1}{2}n(n-1)} \tag{2.2}$$

2.3.5.2 Mesures de qualités basées sur la distance

Sont les plus connus et les plus utilisés pour évaluer la qualité d'une classification. Plusieurs mesures de qualité qui utilisent la distance entre les individus ont été développés dont l'indice de Dunn et la Silhouette.

L'indice de Dunn :

L'indice de Dunn est basé sur l'identification de clusters compacts et bien séparés. Soit dmin la distance minimale entre deux objets de deux différentes classes et dmax la distance maximale entre deux objets de la même classe. Alors, l'indice de Dunn est défini par : D = dmin /dmax.

L'objectif principal de cet indice est de chercher la distance minimale qui sépare deux classes dans la partition tout en tenant compte de la distribution des éléments à l'intérieur des classes. [34] **L'indice de Silhouette :**

Il travaille à l'échelle microscopique, c'est à dire qu'il s'intéresse aux documents en particulier et non pas aux classes. Le but de Silhouette est de vérifier si chaque document a été bien classé. Pour chaque document i de la partition.

on calcule la valeur suivante :

$$-1 \leq S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \leq 1 \tag{2.3}$$

a(i) : représente la distance moyenne qui le sépare des autres documents de la classe à laquelle il appartient.

b(i) : représente la distance moyenne qui le sépare des documents appartenant à la classe la plus proche.

Quand **S(i)** est proche de 1, le document est bien classé : la distance qui le sépare de la classe la plus proche est très supérieure à celle qui le sépare de sa classe. Par contre, si **S(i)** est proche de -1, cela veut dire que le document est mal classé. Mais si **S(i)** est proche de 0 alors il pourrait également être classé dans la classe la plus proche.[34]

2.4 Clustering de documents textuels

2.4.1 Intérêt de clustering des documents textuels

L'intérêt principal du clustering dans les documents textuels est de regrouper des documents similaires ensemble en fonction de leur contenu. Cela permet de mieux organiser et de mieux comprendre de grands ensembles de données de documents textuels, en identifiant des modèles et des structures qui pourraient être difficiles à détecter autrement cette technique présente autres intérêts notamment :

Exploration de données : L'exploration de données est une étape clé dans le processus de clustering de documents textuels. Elle consiste à analyser et à comprendre les caractéristiques des données afin de mieux les structurer et de faciliter le processus de clustering.

L'exploration de données peut inclure des étapes telles que la visualisation des données, l'extraction des caractéristiques, l'analyse des statistiques descriptives, etc. Elle permet d'identifier les modèles, les tendances et les relations entre les documents et de choisir les caractéristiques les plus pertinentes pour effectuer le clustering.

Pour explorer les données dans le contexte du clustering de documents textuels, il est courant d'utiliser des techniques telles que l'analyse de fréquence de termes, qui consiste à identifier les mots clés les plus fréquents dans les documents, ou l'analyse de similarité de documents, qui mesure la similarité entre les documents en fonction de leurs caractéristiques communes.

L'exploration de données peut également aider à identifier les documents qui sont très différents des autres, ce qui peut être utile pour mieux comprendre les données et pour évaluer l'efficacité des méthodes de clustering.

En somme, l'exploration de données est une étape importante dans le processus de clustering de documents textuels, car elle permet de mieux comprendre les données, de choisir les caractéristiques les plus pertinentes pour effectuer le clustering, et d'évaluer l'efficacité des méthodes utilisées.[34]

Résumé automatique : Le résumé automatique peut être utilisé dans le contexte du clustering de documents textuels pour faciliter le processus de catégorisation. En effet, le résumé automatique permet d'extraire les informations clés des documents et de les utiliser comme caractéristiques pour mesurer la similarité entre les documents.

En pratique, le résumé automatique peut être utilisé pour créer des représentations plus courtes des documents, qui peuvent être ensuite utilisées pour effectuer le clustering. En utilisant des techniques de résumé automatique, on peut extraire les concepts et les thèmes les plus importants des documents, et les utiliser comme caractéristiques pour mesurer la similarité entre les documents.

Cela peut aider à réduire la dimensionnalité de l'espace de caractéristiques et à améliorer la précision du clustering. De plus, le résumé automatique peut aider à identifier rapidement les documents les plus pertinents pour une tâche donnée, ce qui peut être utile dans de nombreuses applications, telles que la recommandation de contenu, la veille stratégique, l'analyse de sentiments, etc.[34]

Catégorisation des documents : La catégorisation des documents dans le clustering est un processus par lequel des documents sont regroupés en clusters ou groupes en fonction de leurs caractéristiques communes. Les algorithmes de clustering fonctionnent en mesurant la similarité entre les documents et en regroupant ceux qui présentent des similitudes dans un même cluster. Pour catégoriser des documents dans le clustering, il est important de sélectionner les caractéristiques ou les attributs les plus pertinents qui permettront de mesurer la similarité entre les documents. Cela peut inclure des mots clés, des thèmes, des concepts ou des caractéristiques spécifiques du texte, telles que la longueur ou la fréquence des phrases.

Une fois que les caractéristiques pertinentes ont été sélectionnées, il est possible d'utiliser un algorithme de clustering pour regrouper les documents en clusters. Les résultats du clustering peuvent être analysés pour identifier les tendances ou les modèles dans les données, pour évaluer l'efficacité des caractéristiques sélectionnées et pour aider à la prise de décision dans diverses applications, telles que l'analyse de texte, la recommandation de contenu et la détection de spam.[34]

Recommandations personnalisées : Les recommandations personnalisées sont souvent utilisées dans le clustering de documents textuels pour offrir une expérience personnalisée à l'utilisateur en lui proposant des documents pertinents en fonction de ses préférences et de ses intérêts.

Dans ce contexte, les méthodes de clustering peuvent être utilisées pour créer des groupes de documents similaires, qui sont ensuite utilisés pour recommander des documents pertinents à l'utilisateur en fonction de ses préférences et de ses interactions

précédentes.

Pour réaliser des recommandations personnalisées dans le clustering de documents textuels, on peut utiliser des techniques telles que la classification basée sur les préférences de l'utilisateur, l'apprentissage automatique et le traitement du langage naturel.

En somme, la recommandation personnalisées est une approche importante dans le clustering de documents textuels, car elles permettent d'offrir une expérience utilisateur personnalisée et d'optimiser la pertinence des recommandations en fonction des préférences et des intérêts de l'utilisateur.[34]

Détection de tendances et de sujet : La détection de tendances et de sujets dans le clustering de documents textuels est une méthode pour identifier les sujets les plus pertinents et les tendances actuelles à partir de grands volumes de données textuelles.

Pour y arriver, on peut utiliser des techniques d'analyse de texte, telles que l'analyse sémantique, l'analyse de fréquence de termes et l'analyse de cooccurrences de mots. Ces techniques permettent de détecter les tendances et les sujets les plus fréquents dans les documents textuels, en identifiant les mots clés et les thèmes les plus couramment utilisés.

Une fois que les tendances et les sujets ont été identifiés, on peut utiliser des méthodes de clustering pour regrouper les documents qui traitent de sujets similaires et pour identifier les documents les plus pertinents pour chaque sujet.

En somme, la détection de tendances et de sujets dans le clustering de documents textuels permet d'identifier les sujets les plus pertinents et les tendances actuelles à partir de grands volumes de données textuelles, ce qui peut être utile pour les entreprises, les organisations et les chercheurs qui souhaitent comprendre les tendances actuelles et les sujets les plus importants dans leur domaine d'activité.[34]

2.4.2 Processus de Clustering des documents

Le clustering des documents passe par plusieurs ´étapes illustrés dans cette figure :

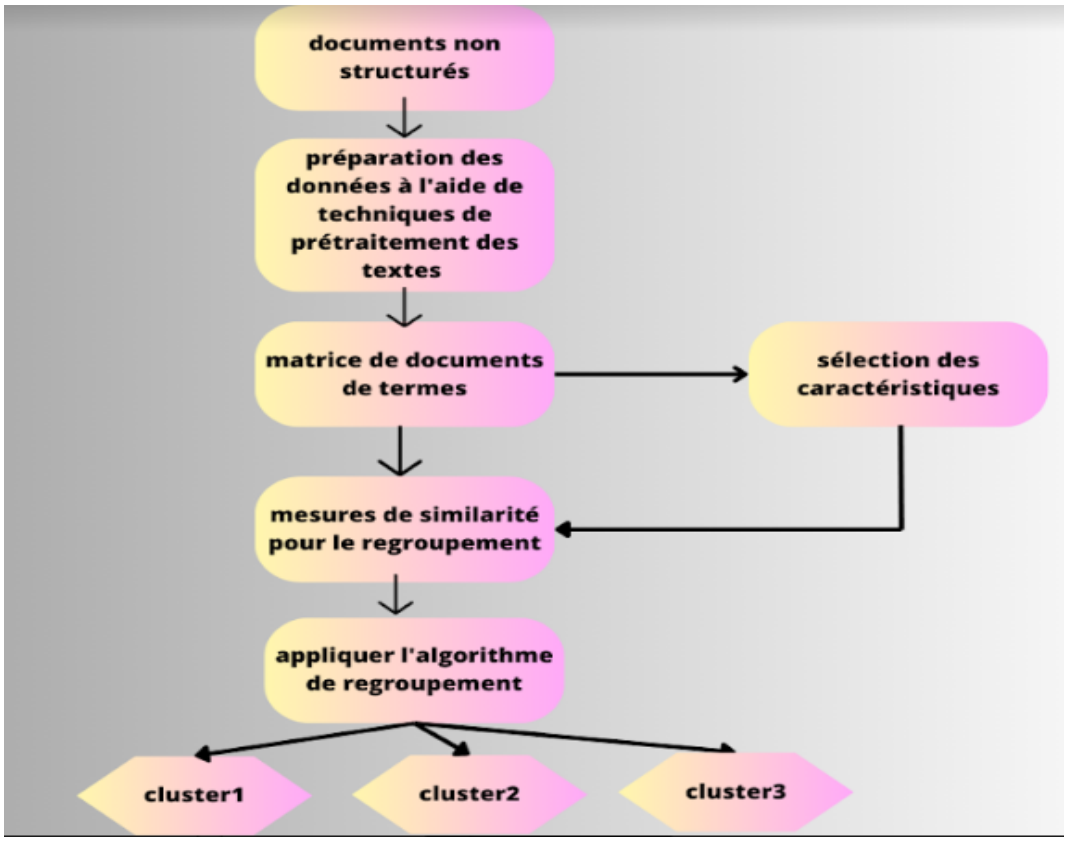

FIGURE 2.9 – Le processus de Clustering

2.4.2.1 Collecte des données

La première étape consiste à collecter les données textuelles à partir de différentes sources, telles que des sites web, des documents, des articles de presse, etc.[36]

2.4.2.2 Prétraitement des données

Le prétraitement des données est une étape clé du traitement des données textuelles, qui vise à nettoyer, normaliser et préparer les données brutes pour une analyse plus approfondie. Le prétraitement des données est une étape importante et nécessaire pour garantir la qualité et la pertinence des résultats de l'analyse de texte. Les étapes de prétraitement peuvent varier en fonction du type de données et des objectifs de l'analyse.[36]

Voici quelques étapes courantes de prétraitement des données pour travailler avec des données textuelles :

Nettoyage des données : Cette étape consiste à éliminer les données bruitées ou inutiles, telles que les balises HTML, les caractères spéciaux, les URL, les emojis, etc. [37]

Normalisation de la casse : Cette étape consiste à uniformiser la casse des données

en minuscules ou majuscules pour éviter les erreurs de traitement dues à des différences de casse.

Tokenisation : La tokenisation des données textuelles est une étape importante dans le prétraitement des données textuelles, qui consiste à découper les données en unités discrètes appelées tokens, comme des mots, des phrases ou des paragraphes. La tokenisation est une étape de segmentation qui permet de transformer les données textuelles en une séquence de tokens qui peuvent être traités par des algorithmes de traitement de texte. Les tokens peuvent être utilisés pour représenter les données textuelles sous forme de vecteurs numériques, qui peuvent être traités par des algorithmes de clustering. La tokenisation peut être effectuée en utilisant des outils de traitement de texte ou de programmation, qui permettent de découper les données textuelles en unités discrètes. La tokenisation est une étape importante pour garantir la qualité des données et améliorer la précision des résultats de l'analyse textuelle.[37]

Suppression des mots vides : Cette étape consiste à éliminer les mots qui n'apportent pas de sens à la phrase, comme les pronoms, les prépositions, les conjonctions, etc.[37]

Lemmatisation ou racinisation : Cette étape consiste à réduire les mots à leur forme de base pour éviter la duplication des informations et pour faciliter la recherche, en utilisant des techniques comme la lemmatisation (réduction à la forme canonique) ou la racinisation (réduction à la racine du mot).[37]

Élimination des doublons : L'élimination des doublons est une étape importante dans le traitement de données textuelles, qui consiste à éliminer les doublons dans les données pour éviter la redondance et les erreurs de traitement. Les doublons peuvent apparaître dans les données textuelles pour différentes raisons, telles que des erreurs de saisie, des copies multiples, etc. L'élimination des doublons permet de réduire la dimensionnalité des données textuelles et d'améliorer la précision des résultats de l'analyse textuelle. [37]

2.4.2.3 Pondération des contenus des documents

La pondération est au coeur de l'étape de génération la matrice terme/documents C'est une étape qui permet d'assigner un poids pour chaque terme dans un document. Ce poids désigne le nombre de fois qu'un certain descripteur (terme) est apparu dans chacun des documents d'un corpus. A partir de ce nombre on peut dire si un descripteur

est discriminant ou non par rapport à un document donné.[37]

Cette étape permet aussi de caractériser les termes importants dans un document, l'idée est que les termes importants doivent avoir un poids important.

Il existe plusieurs techniques de pondération comme le montre le tableau ci-dessous :

Technique	Definition	Forme
Pondération binaire	Comptabiliser la présence de chaque terme dans le document, sans se préoccuper du nombre d'occurrences (de la répétition- TF booléenne)	TF = 1 ou 0
Fréquence des termes-TF (Term fréquency)	TF désigne la fréquence d'un terme (descripteur) dans un texte donné	TF absolu : TF= NT NT est le nombre de fois ou‘ le terme est apparu dans le texte. TF relative : TF =NT/ST NT est le nombre de fois que le terme est apparu dans le document. et, ST est le nombre de tous les termes du document.
Pondération TF-IDF	Relativiser l'importance d'un terme dans un document (TF) par son importance dans le corpus (IDF).	TF-IDF(T,d,D)= TF(T,d)*IDF(T, D) D : document quelconque (TF relative ou absolue)
Fréquence documents inverses (IDF)	Elle mesure le degré de rareté d'un terme, non pas dans un document, mais dans tous les documents (l'influence d'un terme)	IDF(T,D)=$\log_{10} \frac{N}{T}$ D : corpus N : nombre de documents dans le corpus NT : nombre de documents ou‘ le terme apparait

TABLE 2.1 – Les techniques de pondération .

2.4.2.4 Sélection des attributs

Afin d'améliorer l'efficacité et la précision de la classification automatique ,on utilise la technique de sélection des sous-ensemble d'attributs

La sélection des attributs consiste à trouver des sous-ensemble des termes les plus pertinent (important) en utilisant des techniques comme féequence de document (FD), le gain d'information (GI). Ces techniques vont calculer un score pour chaque terme qui servira comme indice de pertinence. Les termes seront donc triées en ordre décroissant pour le but de choisir les mots les plus pertinents selon des critères prédéfinis.[38]

Cette étape permet de régler des problèmes de grande dimensionnalité de la matrice document-terme, ci-dessous on va présenter quelques techniques des élection des attributs :

I. La sélection des attributs basée sur la fréquence de documesnts(FD) :

Le principe de cette méthode est de calculer le nombre de fois qu'un certain terme est apparu dans chacun des documents d'un corpus, cette technique va filtrer les termes ayant une fréquence inférieure à un seuil prédéterminé. Le but de cette technique c'est d'éliminer tous les termes inutiles qui n'ont pas d'influence sur la classification des documents.

II. La sélection des attributs basée sur le gain d'information :

Cette méthode vise à faire la réduction du vocabulaire en se basant sur la présence ou l'absence d'un terme dans un document .Cette technique calcule la valeur de l'entropie (quantité d'information pour chaque document) .c-a'-d il existe une corrélation entre la valeur de l'entropie et la variété des informations dans un document .[38] Le gain d'information d'un terme T dans un ensemble de document D se calcule comme suit :

GAIN(D,T) = Entropie(D) Entropie(T)

Avec :

$$Entropie(D) = \sum_{i=1}^{c} -di \log_2 -di \tag{2.4}$$

Soit c le nombre de catégories dans lesquelles les documents seront classés .

di est le nombre de documents dans D appartenant à la catégorie i .

Et :

$$Entropie(T) = \sum_{v \in 0,1} \frac{|Dv|}{|D|}.Entropie(Dv) \tag{2.5}$$

Soit v la valeur de présence (v=1) ou d'absence (v=0) du terme T dans les documents.

$|Dv|$ le nombre de documents dans D à qui le terme T appartient si (v=1) , ou

n'appartient pas si (v=0) . et l'entropie de Dv est égale à :

$$Entropie(Dv) = \sum_{i=1}^{c} -dvi \log_2 -dvi \tag{2.6}$$

Ou $|dvi|$ est le nombre de documents dans D appartenant à la catégorie i où la valeur du terme T est égale à (0 ou 1) .

2.4.2.5 Choix d'une mesure de similarités

Le choix de la mesure de similarité est une étape importante dans la mise en place d'un algorithme de clustering de documents. Il est essentiel de choisir une mesure de similarité adaptée aux données et aux objectifs de clustering. Plusieurs mesures de similarité sont couramment utilisées pour le clustering de documents.[39]

2.4.2.6 Application d'un algorithme de classification

L'application d'un algorithme de classification implique de catégoriser les données non étiquetées en groupes homogènes, en utilisant des techniques de classification non supervisées, un algorithme est choisi en fonction de la nature des données et de l'objectif de la classification telles que l'algorithme k-means, l'algorithme de clustering hiérarchique, etc.

2.4.2.7 Mesure de la qualité du clustering

La mesure de la qualité du clustering est une étape importante dans le processus de clustering pour plusieurs raisons :

Évaluation objective : Les mesures de qualité fournissent une évaluation objective des résultats du clustering. Elles permettent de quantifier la performance du clustering de manière quantitative, plutôt que de se fier uniquement à une évaluation subjective basée sur une analyse visuelle des clusters.

Comparaison des algorithmes : Les mesures de qualité du clustering permettent de comparer différents algorithmes de clustering ou différentes configurations d'un même algorithme. Elles aident à déterminer quelle méthode ou quel paramétrage donne les meilleurs résultats en termes de qualité du clustering.

Sélection des meilleurs résultats : Les mesures de qualité permettent de sélectionner les meilleurs résultats de clustering parmi plusieurs essais ou exécutions d'un algorithme.

Elles aident à identifier les résultats les plus pertinents et les plus cohérents, ce qui facilite l'interprétation des clusters.

Évaluation de la stabilité du clustering : Certaines mesures de qualité du clustering permettent d'évaluer la stabilité des clusters. Elles mesurent la cohérence des clusters obtenus sur différents échantillons de données ou en utilisant différentes partitions des données. Une meilleure stabilité indique une meilleure qualité et une plus grande robustesse du clustering.

Validation du clustering : Les mesures de qualité du clustering permettent de valider les résultats obtenus. Elles fournissent une indication de la fiabilité et de la pertinence des clusters identifiés.

2.4.3 Mesures de similarité utilisées dans le cadre du clustering des documents

Plusieurs mesures de similarité peuvent être utilisées pour évaluer la similarité entre les documents. Voici quelques-unes des mesures de similarité couramment utilisées :

2.4.3.1 Similarité cosinus

La similarité cosinus est une mesure de similarité couramment utilisée dans le clustering de documents pour comparer la similitude entre deux vecteurs de termes. Elle mesure l'angle entre les deux vecteurs et renvoie une valeur de similarité comprise entre 0 et 1. Plus la valeur de similarité est proche de 1, plus les vecteurs sont similaires, tandis qu'une valeur proche de 0 indique une dissimilitude élevée entre les vecteurs.

La similarité cosinus est souvent utilisée pour la comparaison de textes, car elle permet de prendre en compte la similarité sémantique des termes. Elle est particulièrement utile dans les cas où les documents ont une longueur variable et où les termes peuvent apparaître dans un ordre différent.[40]

La formule de cette méthode est comme suit :

$$cos(u, v) = \frac{u.v}{\|u\| \, \|v\|} \tag{2.7}$$

2.4.3.2 La distance euclidienne

La distance euclidienne est une mesure de distance couramment utilisée en clustering de documents pour mesurer la dissimilarité entre deux vecteurs de termes. Elle est basée sur la géométrie euclidienne et mesure la distance euclidienne entre deux points dans un espace à n dimensions. Dans le clustering de documents, chaque document est représenté sous forme d'un vecteur de termes, où chaque dimension représente un terme et sa valeur représente l'importance ou la fréquence de ce terme dans le document. La distance euclidienne mesure la distance entre deux vecteurs de termes en calculant la racine carrée de la somme des carrés des différences entre chaque dimension des deux vecteurs.[41]

$$d(u,v) = \sqrt{\sum_{j=1}^{p} (uj - vj)^2} \tag{2.8}$$

p est le nombre de termes . uj (vj) est la pondération du terme j pour le document u (resp.v)

2.4.3.3 La similarité de Jaccard

La similarité de Jaccard est une mesure de similarité souvent utilisée en clustering de documents pour mesurer la similarité entre deux ensembles de termes. Elle est basée sur la théorie des ensembles et mesure la proportion d'éléments communs entre deux ensembles.

Dans le clustering de documents, chaque document est représenté sous forme d'un ensemble de termes, où chaque élément représente un terme présent dans le document. La similarité de Jaccard mesure la similarité entre deux ensembles de termes en calculant le nombre d'éléments communs divisé par le nombre total d'éléments dans les deux ensembles.[42]

$$J(A,B) = \frac{|A \cap B|}{A \cup B} = \frac{A \cap B}{|A| + |B| - A \cap B} \tag{2.9}$$

2.4.3.4 Indice de Dice

Mesurer la similarit´e entre deux documents en se basant sur le nombre des termes communs entre eux.[43]

$$d(u,v) = \frac{2Nc}{N1 + N2} \tag{2.10}$$

Nc est le nombre des termes communs entre les documents u(resp.v) N1 (resp. N2) le nombre des termes de u(resp.v).

2.5 Conclusion

Dans ce deuxième chapitre, nous avions commencé par aborder le domaine de l'apprentissage automatique (Machine Learning). Nous avions ensuite présenté la technique du Clustering qui est une technique d'apprentissage automatique non supervisé. Nous avons présenté son principe de fonctionnement, ces algorithmes les plus connus ainsi que les mesures de son évaluation. En troisième lieu, nous avions abordé le Clustering des documents textuels et ses spécificités étant donné que le but de notre travail c'est d'appliquer le clustering dans les documents scientifiques.

Le prochain chapitre sera dédié justement à présenter les éléments de notre solutions dans le cadre du clustering automatiques des documents scientifiques qui sont caractérisés par la présence d'autres information en plus du contenu textuels comme les information de structure ainsi que les liens de citation qui peuvent être exploité pour mieux les organiser et les regrouper.

Chapitre 3

Système de Clustering des documents scientifiques

3.1 Introduction

Avec la croissance exponentielle du nombre d'articles scientifiques disponibles au format numérique, il devient de plus en plus difficile de gérer cette masse considérable d'informations. Il est essentiel de concevoir des systèmes automatisés de classification des articles afin d'organiser efficacement cet espace de données et de faciliter l'accès à l'information pertinente.

Nous rappelons ici que l'objectif principal de notre travail est de développer un système de classification automatique des articles scientifiques qui :

- Offre une vue globale des principaux thèmes de recherche abordés dans une collections d'articles scientifiques et facilite l'exploration de l'espace scientifique.

- Permet à l'utilisateur de trouver rapidement les articles en rapports avec son domaine d'intérêt,

- Détecter automatiquement les différents sujets et problématiques de recherche abordé dans une grande collection de papiers scientifique .

Avant de présenter les différentes étapes du développement de notre système, nous aborderons les défis de cette problématique, les possibilités offertes par les caractéristiques des documents scientifiques qui peuvent être exploitées dans le cadre d'un clustering. Nous présentons aussi quelques travaux connexes dans ce domaine qui montre comment on peut exploiter les différentes composantes comme le contenu, la structure, les métadonnées et

les réseaux de citations dans un processus de Clusterings d'articles scientifiques.

3.2 Défis liés à la classification des articles

La classification des articles scientifiques présente plusieurs défis importants auxquels il est nécessaire de faire face. Voici quelques-uns des défis les plus couramment rencontrés :

- **Volume et variété des données :** Avec le nombre croissant d'articles scientifiques disponibles, la quantité de données à traiter peut-être énorme. De plus, les articles peuvent provenir de divers domaines scientifiques, ce qui rend la classification encore plus complexe.

- **Hétérogénéité des sources d'information :** Les articles scientifiques peuvent provenir de différentes sources, telles que des revues spécialisées, des conférences, des archives en ligne, etc. Chaque source peut avoir ses propres normes de publication et de formatage, ce qui ajoute une couche de complexité à la classification.

- **Évolution des domaines scientifiques :** Les domaines scientifiques évoluent rapidement, de nouveaux concepts émergent et les frontières entre les disciplines peuvent devenir floues. La classification des articles scientifiques doit être en mesure de s'adapter à ces évolutions et de prendre en compte les tendances émergentes.

- **Manque d'étiquetage ou d'annotations :** Dans de nombreux cas, les articles scientifiques ne sont pas préalablement étiquetés ou annotés avec des informations de classification. Cela signifie que des techniques d'apprentissage automatique non supervisées ou semi-supervisées doivent être utilisées pour découvrir les structures et les relations dans les données.

- **Grande taille de l'espace de représentation :** Etant donné le format textuel des articles scientifique, la représentation de chaque document se fait souvent sous forme d'un vecteur pondéré de termes. Le nombre de dimension de l'espace de représentation est de ce fait très grand. Même si des techniques de réduction de cette espace telles que l'élimination des mots outils ou l'élimination des termes les moins fréquents, la taille de l'espace de représentation reste grande.

Pour relever ces défis, des approches avancées utilisant des techniques de traitement automatique du langage naturel, d'apprentissage automatique et de fouille de données sont souvent nécessaires. Il est également important de combiner différentes sources d'information, telles que les citations, le contenu textuel, les métadonnées, etc., pour obtenir

une classification plus précise et complète des articles scientifiques.

3.3 Travaux dans le domaine

Cette section fournit un aperçu des recherches antérieures pertinentes et explique comment le travail de regroupement des documents scientifiques basé sur la similarité des citations et du texte s'inscrit dans le paysage de la recherche existant. En général les travaux peuvent être classées en deux catégories :

3.3.1 L'approche de similarité basée sur le texte

Ceux qui se sont basé sur le contenu textuel des documents pour réaliser le clustering. Ici les documents sont comparés sur la base de leurs contenu textuel. Deux documents sont considérés dans le même cluster si la similarité entre leurs contenus textuel est forte.

3.3.2 L'approche de similarité basée sur les citations

Ceux qui exploitent les réseaux de citations pour réaliser le clustering. Ici les documents sont comparés sur la base de leurs bibliographie. Deux documents sont considérés dans le même cluster si la similarité entre leurs bibliographie est forte.

3.3.3 L'approche de similarité hybride basée sur le texte et les citations

Il faut noter aussi que dernièrement, certains travaux ont essayé de proposer une approche hybride ou les deux composantes : textuelle et de citation sont considérées dans le processus de clustering

Parmi les travaux de l'approche text-based similarity, nous pouvons citer :

• [44] où les auteurs ont développé dans leurs article (Scientific Documents Clustering Based on Text Summarization) une solution de clustering de documents scientifiques basé sur le résumé de texte pour le regroupement de documents. Elle comprend une phase de prétraitement où les étapes classiques de tokenisation, lemmatisation et élimination des mot vides sont appliquées réduisant ainsi la quantité de données à traiter. Ensuite, une phase de pondération via le calcul des scores TF-IDF et BM25 sont calculés pour chaque

mot. Enfin, Une phase de résumé basé sur les scores BM25 de l'étape précédente permet de ne garder de chaque document que les phrases les plus pertinentes et qui permet de réduire encore la taille des données à traiter dans la phase de Clustering. Cette solution a donné de très bons résultats dans l'évaluation surtout en termes d'efficacité et du temps d'exécution.

• Dans [45] les auteurs ont développé une approche de clustering basée sur les passages à l'intérieur des documents (Text Document Clustering on the basis of Inter passage approach by using K-means). Dans leur approche, ils considèrent qu'un même document peut être concernés par plusieurs sujets ou topic, et donc ils effectuent le clustering au niveau intra-document (niveau passages ou segments de document). Une fois les segments dans chaque document sont classés. Ensuite un clustering inter-document est effectué ou les segments de toute la collection sont classés. TF-IDF et le score score SentiWordNet pour identifier les mots-clés importants associés à chaque segment de doucement. Une fois les mots-clés identifiés pour chaque segment, ils ont choisi les mot-clé ayant le score le plus élevé, pour ce segment particulier. Ensuite, ils ont calculé le score global du segment. En utilisant ces scores de segment, ils ont appliqué l'algorithme K-means pour le regroupement inter-documents.

Dans l'approche citation-based similarity, on peut citer :

• [46] une méthode a été introduite pour estimer la similarité entre les articles scientifiques en utilisant un moteur de recherche sur le Web ; est appelée la méthode des co-citations sur le Web. Dans cette méthode, la similarité entre deux articles est basée sur le nombre de fois où ils ont été mentionnés ensemble sur le Web ; la similarité entre deux articles est basée sur leur nombre de co-références , Il a supposé que si deux articles aient des références communes, ils ont probablement un même sujet. Ils ont calculé la similarité de co-citation entre deux articles scientifiques par le nombre de fois où ils sont co-cités sur le Web en utilisant le moteur de recherche Google. Des expérimentations ont été réalisées pour comparer la performance de différentes méthodes de citations : le couplage bibliographique (bibliographique coupling) , la co-citation traditionnelle avec la base de données de citation Web of Science.

• [2] : cette étude se concentre sur le regroupement des documents scientifiques basé sur le modèle de citation étendu qui considère la fréquence et la distribution des documents scientifiques cités dans d'autres documents, la similarité entre les ces documents basés sur

le modèle proposé est composée de trois réseaux : réseau de citation directe, réseau de co-citation, réseau couplage bibliographique, ils ont utilisé l'algorithme K-means pour le regroupement

Dans les approches hybrides :

Les auteurs dans [47] étendent un modèle de clustering hybride en introduisant la notion de "documents centraux" (core documents). Pour ce faire, ils construisent d'abord un réseau appelé le réseau Amsler, qui intègre à la fois les liens de citation et les liens de co-citation. Ce réseau est utilisé pour calculer la similarité basée sur les citations en utilisant l'angle cosinus entre les articles. Ensuite, la similarité basée sur le texte est également calculée en utilisant la similarité cosinus, qui prend en compte à la fois les caractéristiques textuelles statistiques et topologiques des articles.

3.4 Système de Clustering de documents scientifiques

3.4.1 Considérations de base de notre solution

Avant de détailler le fonctionnement global de notre solution, nous présentons ici les éléments de réflexion sur lesquels nous nous somme appuyer pour son développement. Ces éléments sont comme suit :

1. Exploiter les spécificités des documents scientifique dans le processus de Clustering.
2. Combiner plusieurs facteurs de similarité (textuelle, à base de citation,...etc).
3. Prendre en considération la fréquence des citations dans le calcul de la similarité à base de citation. Nous considérons que le nombre de fois qu'une citation apparait dans un document doit aussi être considéré dans le calcul des similarités.

3.4.1.1 Exploiter les spécificités des articles scientifique dans le processus du clustering

Comme mentionné dans le chapitre 2, Un article scientifique est un document académique qui a des spécificités propres qui le différencie des documents autres textuels non structuré.

o En plus de contenu textuel brute, un article scientifique à une structure rigoureuse (title, abstract, key-words, introduction, methods, results, discission, references ...etc).

o Les documents scientifiques se caractérisent aussi par leurs liens de références et de citations qu'il faudrait bien exploiter dans le cadre du Clusetring.

o Un document scientifique possède des métadonnées : auteurs, mot-clé, nom de revue ou conférences qui peuvent aussi être exploité dans le processus de Clustering.

Nous envisageons d'exploiter toutes ces aspects dans le processus du clustering soit en phase de représentation soit en phase de comparaisons entre les documents.

3.4.1.2 Combiner entre plusieurs facteurs de similarités

Dans tout processus de Clustering, on a besoin d'une mesure de similarité pour comparer chaque paire d'éléments. Dans le cadre des documents scientifiques, plusieurs facteurs peuvent être déterminants dans le calcule de similarité entre les documents. Il est évident qu'il faudrait exploiter et combiner ces facteurs pour un meilleure regroupement :

o Tout d'abord la similarité sur la base du contenu textuel brute des documents, mais pas que.

o Les liens de citations entre une paire de documents peuvent aussi donner une indication sur la similarité entre deux documents. Deux documents qui partagent un grand nombre de références sont très probablement proche sémantiquement.

o Les profiles des auteurs peuvent aussi être un indicateur sur la similarité d'une paire d'articles scientifique. Deux articles écrits pat deux auteurs ayant un profils scientifiques proche sont susceptibles d'être aussi proches.

o Le titre de revue, les mots-clés aussi peut être considérée. En plus, il ne faut pas oublier de prendre en considération la qualité du Clustering en termes d'efficacité. Etant donné la taille importante de représentation, il est important de proposer une solution avec une bonne performance du Clustering en termes d'efficacité (temps d'exécution).

3.4.2 Architecture générale du système

La figure suivante montre l'architecture globale de notre système. Nous donnerons les détails de nos propositions à chacune des étapes :3.1

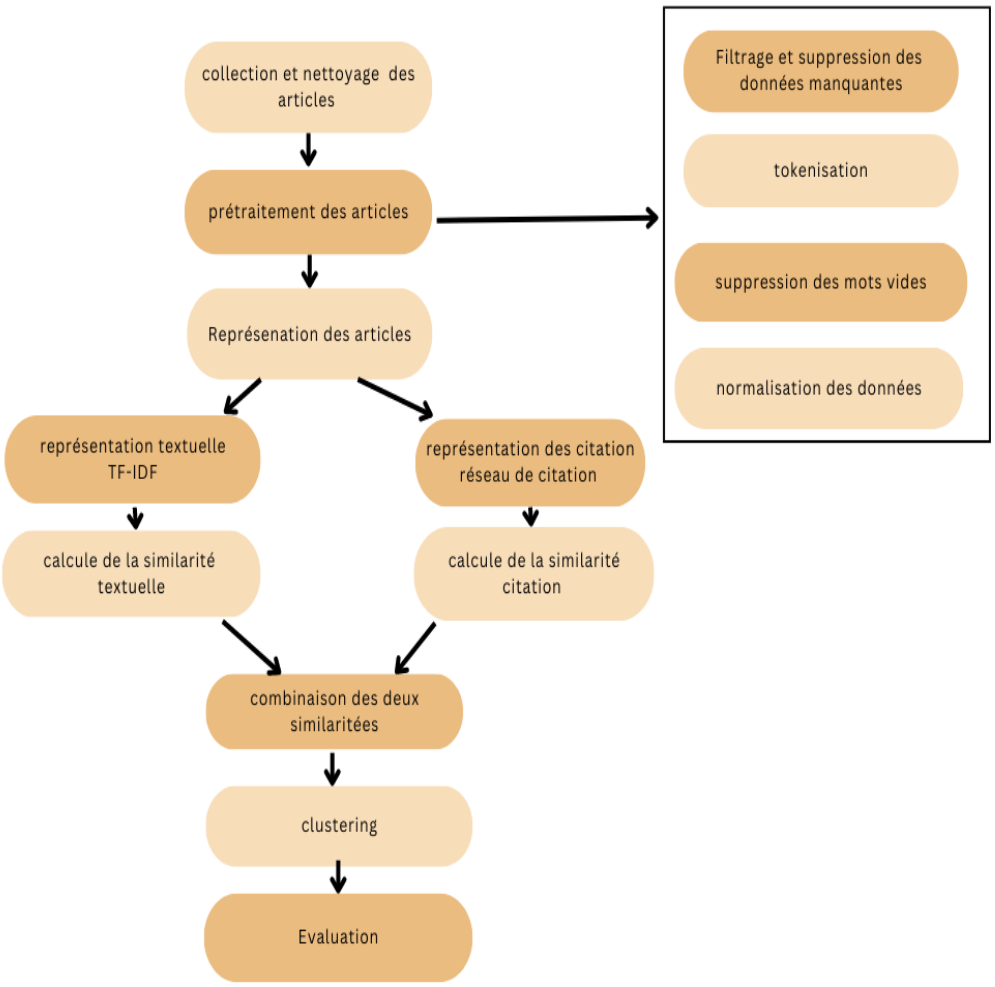

FIGURE 3.1 – Architecture générale du système

3.4.2.1 Prétraitement des données

Le prétraitement des données est une étape standard dans tout processus d'apprentissage automatique. Cette étape comporte les traitements classiques de préparation de données textuelles à savoir :

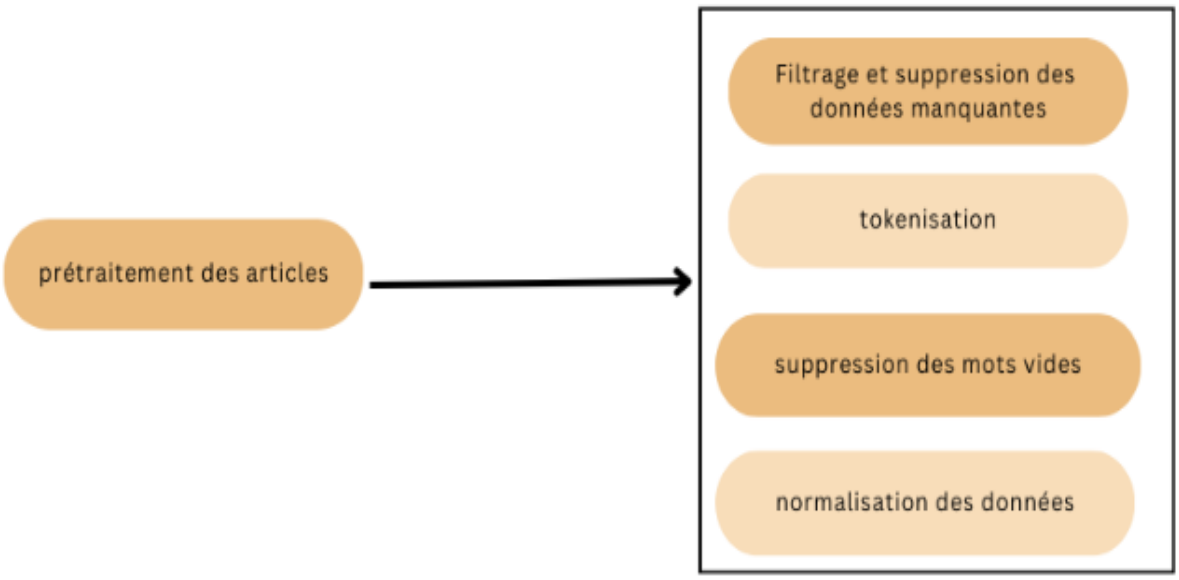

FIGURE 3.2 – Le prétraitement des articles

• **Filtrage et suppression des données manquantes :** Le filtrage des données manquantes consiste à décider comment gérer ces valeurs manquantes de manière appropriée.

• **La tokenisation :**

Elle nous permet de représenter les données textuelles de manière structurée. En divisant le texte en tokens, Cela permet de créer des vecteurs de caractéristiques qui représentent chaque document. Ensuite, ces vecteurs de caractéristiques sont utilisés pour mesurer les similarités sur la base du contenu entre les documents.

Cette figure illustre un exemple de tokenisation :3.3

```
TITRE:
Studies on Cardinality of Solutions for Multilayer Nets and a Scaling Method in Hardware Implementations
Tokens:
['Studies', 'on', 'Cardinality', 'of', 'Solutions', 'for', 'Multilayer', 'Nets', 'and', 'a', 'Scaling', 'Method', 'in', 'Hardwa
re', 'Implementations']
ABSTRACT:
Power systems are evolving to accommodate increased renewable sources of energy, distributed control systems, active distributi
on systems, and consumer-involved energy management systems. This evolution is characterized by the integration of a communicat
ion network overlay that facilitates the bilateral flow of energy and information. In this paper, we present an offline smart g
rid co- simulator test-bed using pre-existing communication and power simulators. A detailed description of the test-bed setup
and implementation is provided in this paper to help facilitate the study of relevant problems by researchers in the field. The
test-bed is targeted as a tool to help researchers verify various communication enabled control schemes designed for distributi
on system, and assess system control resilience against common cyber threats.
Tokens:
['Power', 'systems', 'are', 'evolving', 'to', 'accommodate', 'increased', 'renewable', 'sources', 'of', 'energy', ',', 'distrib
uted', 'control', 'systems', ',', 'active', 'distribution', 'systems', ',', 'and', 'consumer-involved', 'energy', 'management',
'systems', '.', 'This', 'evolution', 'is', 'characterized', 'by', 'the', 'integration', 'of', 'a', 'communication', 'network',
'overlay', 'that', 'facilitates', 'the', 'bilateral', 'flow', 'of', 'energy', 'and', 'information', '.', 'In', 'this', 'paper',
',', 'we', 'present', 'an', 'offline', 'smart', 'grid', 'co-', 'simulator', 'test-bed', 'using', 'pre-existing', 'communicatio
n', 'and', 'power', 'simulators', '.', 'A', 'detailed', 'description', 'of', 'the', 'test-bed', 'setup', 'and', 'implementatio
n', 'is', 'provided', 'in', 'this', 'paper', 'to', 'help', 'facilitate', 'the', 'study', 'of', 'relevant', 'problems', 'by', 'r
esearchers', 'in', 'the', 'field', '.', 'The', 'test-bed', 'is', 'targeted', 'as', 'a', 'tool', 'to', 'help', 'researchers', 'v
erify', 'various', 'communication', 'enabled', 'control', 'schemes', 'designed', 'for', 'distribution', 'system', ',', 'and',
'assess', 'system', 'control', 'resilience', 'against', 'common', 'cyber', 'threats', '.']
```

FIGURE 3.3 – Exemple de tokenisation

• **Suppression des mots vides :** Dans le cadre de notre cas, la suppression des mots vides permet de réduire le bruit et de se concentrer sur les mots plus informatifs et distinctifs. Cela peut conduire à une meilleure représentation des documents et à des résultats de clustering plus précis.

• **Lemmatisation :** La lemmatisation permet de ramener chaque mot à sa racine grammaticale. Cette étape permet aussi de réduire l'espace de représentation. La lemmatisation consiste à analyser les termes de manière à identifier leurs formes canoniques (lemmes) afin de réduire les différentes formes (pluriel, féminin, conjugaison, etc.).

Par exemple le Lemme des mots **" ´economie"** , **´economiquement"** ,**" ´econome"**, et **´econométrie** est le lemme **"ECONOM´** .

Le résultat de lemmatisation du texte suivant :

> « *The continually developing Internet generates a considerable amount of text data When attempting to extract general topics or themes from a massive corpus of documents dealing with such a large volume of text data in an unstructured format is a big problem Text document clustering TDC is a technique for grouping texts based on their content similarity Partitioning text collection based on the documents content significance is one of the most challenging tasks at TDC This study proposes the Bare Bones Based Salp Swarm Algorithm BBSSA to solve the problem of TDC In addition to extract the topics from the clusters an ensemble approach for automatic topic*

extraction TE is proposed ≫.

Donnera ce résultat :

≪ The continu develop Internet gener a consider amount of text data When attempt to extract gener topic or theme from a massiv corpu of document deal with such a larg volum of text data in an unstructur format is a big problem Text document cluster TDC is a techniqu for group text base on their content similar Partit text collect base on the document content signific is on of the most challeng task at TDC Thi studi propos the Bare Bone Base Salp Swarm Algorithm BBSSA to solv the problem of TDC In addit to extract the topic from the cluster an ensembl approach for automat topic extract TE is propos ≫.

3.4.2.2 Représentation des données

La représentation des données consiste à choisir une représentation appropriée qui capture les caractéristiques pertinentes des données pour la tâche de clustering. Comme nous l'avions mentionné plus haut, nous envisageons d'exploiter la richesse du document scientifique dans sa représentation. Dans le cadre de notre travail on a deux aspects de représentation des données :3.4

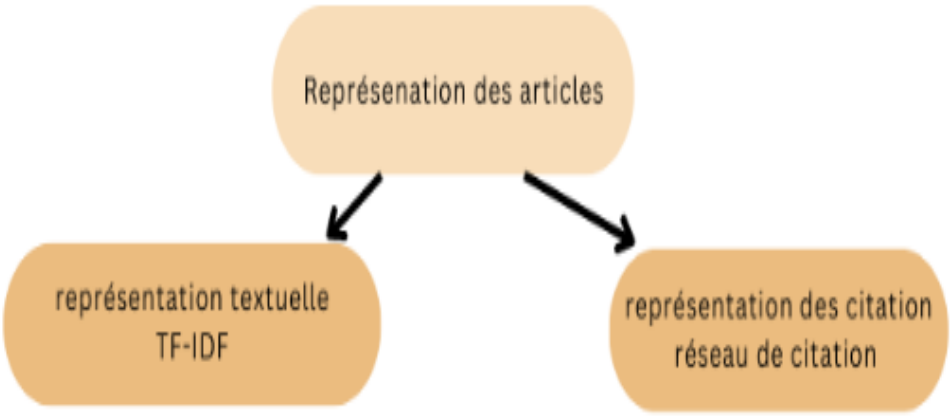

FIGURE 3.4 – La représentation des articles

-Représentation textuelle :

Concernant le contenu textuel, nous avons sélectionné le titre et le résumé de chaque article. Le contenu de ces deux parties représente à notre avis le mieux le sujet abordé dans l'article scientifique. Ce contenu textuel (titre + résumé) passe par les étapes du prétraitement citées ci-dessus (tokenisation, élimination de mots vide puis lemmatisation) pour former un vecteur de tokens pour chaque document.

La représentation textuelle consiste à caractériser l'importance de chaque token dans la représentation du contenu des documents. Cela passe par une phase de pondération qui permet de donner un poids à chaque token dans un document selon son importance. Plusieurs méthodes de pondération sont disponibles (voir chapitre 2). Pour notre cas nous avons opté pour une pondération en **TF-IDF** qui est la plus utilisée dans ce domaine.

Le résultat de cette phase c'est une matrice qui comporte en lignes les articles et en colonnes les différents token. Une **cellule [i,j]** donne le poids **TF-IDF** du terme **j** dans l'article **i**.Comme suit :3.6

	Document	informal	formal	correspond	construction	show	across	consistent	general	differently	...	sport	located	fails	managed
0	1	0.08997	0.081162	0.086458	0.075548	0.034818	0.070252	0.077206	0.057741	0.094498	...	00	00	00	00
1	2	00	00	00	00	00	00	00	00	00	...	00	00	00	00
2	3	00	00	00	00	00	00	00	00	00	...	00	00	00	00
3	4	00	00	00	00	00	00	00	00	00	...	00	00	00	00
4	5	00	00	00	00	0.038811	00	00	00	00	...	00	00	00	00
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
441	442	00	00	00	00	00	00	00	00	00	...	00	00	00	00
442	443	00	00	00	00	00	00	00	00	00	...	00	00	00	00
443	444	00	00	00	00	0.027665	00	00	00	00	...	00	00	00	00
444	445	00	00	00	00	0.039316	00	00	00	00	...	00	00	00	00
445	446	00	00	00	00	00	00	00	00	00	...	0.079649	0.079649	0.079649	0.159299

FIGURE 3.5 – La matrice TF-IDF

Vu l'importance du titre d'un article scientifique dans l'indication de son sujet, dans notre approche, nous avons attribué plus de poids aux titres des documents avant de représenter les données (TF-IDF). Cette étape de pondération préalable permet de mettre en évidence l'importance des mots clés présents dans les titres des documents lors du processus de clustering. En accordant une attention particulière aux titres, nous pouvons capturer les caractéristiques les plus saillantes des documents dès le départ et ainsi améliorer

la représentation des données textuelles. Cela contribue à une meilleure distinction entre les documents et peut conduire à des clusters plus cohérents et informatifs.

-Pour la représentation des citations :

On a utilisé le réseau de citation qui représente les documents scientifiques sous forme de nœuds et les liens de citation entre les documents comme des arêtes dans un réseau de citation. Cette représentation permet de capturer les relations de citation et de détecter les communautés ou les clusters de documents qui partagent des citations similaires.

Cette figure illustre le réseau de citation :

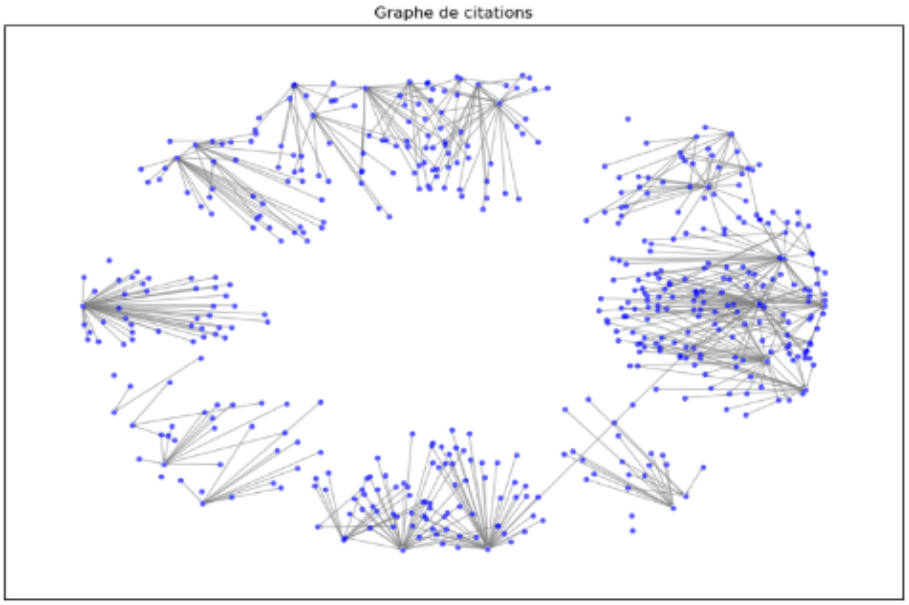

FIGURE 3.6 – Réseau de citation

3.4.2.3 Calcule de la similarité

Notre étude vise à évaluer à la fois la similarité textuelle et la similarité des citations. La similarité textuelle nous permet d'évaluer la proximité sémantique entre les documents en se basant sur leur contenu textuel, tandis que la similarité des citations nous permet de capturer les relations de citation entre les documents.

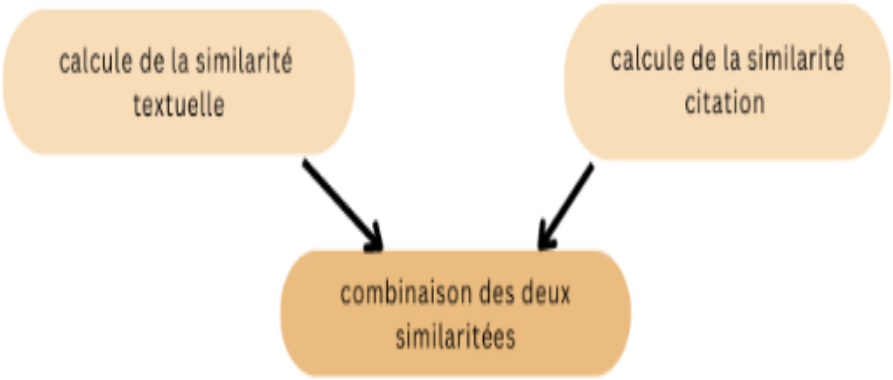

FIGURE 3.7 – Les deux similarités

- Calcule de la similarité textuelle : Pour calculer la similarité textuelle, nous avons utilisé la mesure de similarité cosinus. Cette mesure est basée sur la similarité des vecteurs représentant les documents dans un espace vectoriel. En utilisant la représentation vectorielle des documents obtenue à partir du TF-IDF, nous calculons le produit scalaire entre les vecteurs des documents et mesurons leur similarité à l'aide de la formule du cosinus. Plus la valeur du cosinus est proche de 1, plus les documents sont similaires sur le plan textuel. Cette approche nous permet d'évaluer la proximité sémantique entre les documents. Voici la formule de similarité cosinus :

$$similarity(A, B) = \frac{A.B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} Ai \times Bi}{\sqrt{\sum_{i=1}^{n} A^2 i} \times \sqrt{\sum_{i=1}^{n} B^2 i}} \tag{3.1}$$

- A et B sont deux vecteurs que vous souhaitez comparer.
- (A. B) représente le produit scalaire des vecteurs A et B.
- $\|A\|$ et $\|B\|$ représentent les normes (longueurs) des vecteurs A et B.

La figure suivante représente visuellement la similarité entre deux documents :

Document	1	2	3	4	5	6	7	8	9	10	...	437	438	439	440	
Document																
1	1.000000	0.007611	0.000000	0.001643	0.024586	0.021945	0.003271	0.021854	0.004134	0.028790	...	0.016204	0.011902	0.024632	0.031608	0.02
2	0.007611	1.000000	0.016821	0.007207	0.020839	0.012551	0.029079	0.000000	0.011656	0.043405	...	0.059604	0.026806	0.169686	0.077746	0.03
3	0.000000	0.016821	1.000000	0.032019	0.013423	0.012701	0.000000	0.018950	0.026119	0.047263	...	0.003270	0.005016	0.008099	0.058822	0.02
4	0.001643	0.007207	0.032019	1.000000	0.021325	0.000000	0.107054	0.019120	0.003413	0.026149	...	0.027855	0.000000	0.031742	0.084452	0.01
5	0.024586	0.020839	0.013423	0.021325	1.000000	0.010482	0.041334	0.018314	0.000000	0.060893	...	0.014689	0.019092	0.054074	0.039660	0.00
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
442	0.004379	0.015834	0.031871	0.080323	0.011623	0.004477	0.222773	0.040156	0.006867	0.027533	...	0.006202	0.018828	0.026258	0.079563	0.01
443	0.030947	0.010544	0.011646	0.005633	0.019822	0.015032	0.000000	0.008913	0.008833	0.011821	...	0.016680	0.019060	0.010501	0.009255	0.02
444	0.025626	0.029125	0.000000	0.005842	0.035343	0.029727	0.002961	0.000000	0.001125	0.028578	...	0.021506	0.000000	0.018931	0.039844	0.02
445	0.006487	0.092527	0.036543	0.019833	0.010199	0.036232	0.025834	0.026839	0.001598	0.110062	...	0.013427	0.046942	0.063210	0.085017	0.02
446	0.008103	0.033140	0.006491	0.005952	0.024598	0.020956	0.007475	0.005348	0.015532	0.014807	...	0.028037	0.015808	0.113577	0.070444	0.02

FIGURE 3.8 – La similarité entre deux documents

- La similarité à base de citations :

Comme mentionné dans le chapitre précédent, l'analyse du réseau de citations peut nous fournir des informations très utiles sur les liens entre les thèmes abordés par les articles scientifiques. La similarité sur la base des citations peut être évaluée selon les trois méthodes suivantes :

• **Direct citation similarity :** deux document d1 et d2 sont proche s'il existe un lien de citation directe entre eux (ex d1 cite d2).

La formule classique de calcule de similarité sur la base de Direct citation est la suivante :

$$S^{dt}_{\mathrm{i,j}} = \frac{t_{i,j}}{T_i} + \frac{t_{j,i}}{T_j} \tag{3.2}$$

où :

-ti ;j représente le nombre de fois que le document j est cité dans le document i.

- tj ;i représente le nombre de fois ce document i est cité dans le document j.

-Ti représente le nombre total de citations du document i.

- Tj représente le nombre total de citations du document j .

• **Co-citation similarity :** deux document d1 et d2 sont considéré proche si un autre document dk cite d1 et d2.

La formule classique de calcule de similarité sur la base de Co-citation est la suivante :

$$S^{tco}_{\mathrm{i,j}} = \frac{|Ci \cap Cj|}{|Ci \cup Cj|} \tag{3.3}$$

où :

-Stco i ;j représente la similarité entre les documents i et j.

-Ci et Cj représentent la collection de documents qui citent des documents i et j, respectivement.

• **Bibliographic cooopling similarity :** deux document d1 et d2 sont considérés proche si les deux citent le même document dk. D'après la littérature le bibliographic coopling donne de meilleurs résultats dans le cadre du clustering.

La formule classique de calcule de similarité sur la base de bibliographic coopling est la suivante :

$$S^{tb}_{\mathrm{i,j}} = \frac{|C'i \cap C'j|}{|C'i \cup C'j|} \tag{3.4}$$

Où :

-Stbij représente la similarité entre les documents i et j basée sur le réseau de couplage bibliographique traditionnel.

-C'i et C'j représente l'ensemble des documents qui sont cités respectivement par les documents i et j.

Cette formule de calcule de similarité ne prend pas en considérations la fréquence des citations. C'est-à-dire le nombre de fois qu'une citation apparait dans un document.

Par exemple si d1 et d2 cite un document d3, le fait que d1 le cite 1 fois et d2 le cite 6 fois est considéré de la même manière que si d1 et d2 le citent d3 le même nombre de fois chacun. La similarité entre d1 et d2 sera la même quelques soit le nombre de fois que d1 cite d3 respectivement d2 cite d3.

Notre proposition c'est d'introduire ce facteur de fréquence dans le calcul de cette similarité.

Soit d1 et d2 deux document

Soit C1 : l'ensemble de références de d1 et C2 l'ensemble de références de d2

Nous adoptons la formule suivante pour le calcule de similarité entre d1 et d2 sur la base du bibliographic coopling :

$$Sim(d1, d2) = \frac{\sum_{ci \in C1 \cap C2} \frac{nCid1 + nCid2}{2 * \max(nCid1, nCid2)}}{|C1 \cup C2|} \tag{3.5}$$

Avec

- $C_1 \cap C_2$: l'ensemble des docs qui sont cités par d1 et d2.

- $|C_1 \cup C_2|$: c'est le nombre de références de d1 et d2 ; ici rien n'a changé.

- nCid1 : le nombre de fois ou le doc Ci est référencé dans d1.

- nCid2 : le nombre de fois ou le doc Ci est référencé dans d2.

Prenons un exemple pour illustrer la formule :

Les références de d1 sont : d5,d6, d7, d8, d9.

Les références de d2 sont : d5,d6, d17, d18, d99, d100.

Avec la formule classique :

Sim (d1,d2) = 2/(11) = 0.181

Supposant maintenant que d5 a été cité une fois dans d1 et 8 fois dans d2, et que d6 a été cité 4 fois dans d1 et 5 fois dans d2 La similarité avec la nouvelle formule sera :

Sim(d1,d2) = ((((1+8) / (2*8))) + (((4+5) / (2*5)))) / 11 = 0.133

Alors que si d5 a été cité 5 fois dans d1 et 8 fois dans d2, et que d6 a été cité 4 fois dans d1 et 4 fois dans d2

Sim(d1,d2) = ((((5+8) / (2*5))) + (((4+4) / (2*4)))) / 11 = 0.21

Voici les résultats de la similarité :

```
Similarité entre documents aa9840eb-2cbd-4a38-a55d-11692e651722 et 627e5b37-1246-475d-9715-8bbedb91f807: 0.015384615384615385
Similarité entre documents aa9840eb-2cbd-4a38-a55d-11692e651722 et 19858632-a538-44f4-9062-aa6af347b691: 0.01282051282051282
Similarité entre documents aa9840eb-2cbd-4a38-a55d-11692e651722 et 1257ef32-7c2a-4c45-8cfd-76230197d396: 0.01694915254237288
Similarité entre documents aa9840eb-2cbd-4a38-a55d-11692e651722 et 27966028-0876-421c-9ee0-3593e55436ab: 0.030303030303030304
Similarité entre documents aa9840eb-2cbd-4a38-a55d-11692e651722 et db485b8f-288a-40e5-8e9f-42d0d596c185: 0.029850746268656716
Similarité entre documents aa9840eb-2cbd-4a38-a55d-11692e651722 et cf15ed3c-defd-4dbd-9179-5c57b983a5f0: 0.034482758620689655
Similarité entre documents aa9840eb-2cbd-4a38-a55d-11692e651722 et 1fc103f3-d6bd-4356-aee8-e698b5b71f3a: 0.03225806451612903
Similarité entre documents aa9840eb-2cbd-4a38-a55d-11692e651722 et a32b0bda-eb0c-48d1-b584-43ff8243d4e6: 0.05263157894736842
Similarité entre documents aa9840eb-2cbd-4a38-a55d-11692e651722 et 9abe2ed8-39f4-47ca-8c55-70cab0392658: 0.058823529411764705
Similarité entre documents bd6d40eb-4849-48d3-b739-2df30078b188 et 2eaccba0-96ec-4b54-a7ae-f75003142bad: 0.020833333333333332
Similarité entre documents bd6d40eb-4849-48d3-b739-2df30078b188 et ece79570-1bf1-4b62-bb59-737bc1487e4e: 0.018867924528301886
Similarité entre documents 42439658-0c8f-4ce7-a9a2-004b514f782f et b8724cf1-5ab4-44e9-bbfb-e178d524fa14: 0.02702702702702703
Similarité entre documents db6b59b9-9049-4932-bd55-763229ee1214 et 30841dad-a676-431e-8ef0-c07839c1106e: 0.015151515151515152
Similarité entre documents db6b59b9-9049-4932-bd55-763229ee1214 et ba5f6df8-b11c-4c04-a462-4473c44f51fa: 0.02040816326530612
Similarité entre documents 7a0bad1a-3e0c-48d0-b62b-b726422a904b et 19858632-a538-44f4-9062-aa6af347b691: 0.014084507042253521
Similarité entre documents 7a0bad1a-3e0c-48d0-b62b-b726422a904b et 1257ef32-7c2a-4c45-8cfd-76230197d396: 0.019230769230769232
```

FIGURE 3.9 – Exemple de similarité des citations

3.4.2.4 Combinaison des deux similarités

La combinaison des deux similarités, à savoir la similarité textuelle et la similarité des citations, nous permet d'obtenir une vue d'ensemble plus complète des relations entre les documents scientifiques. En combinant ces deux aspects, nous sommes en mesure

d'explorer à la fois la proximité sémantique basée sur le contenu textuel et les liens de citation entre les documents.

Pour obtenir une combinaison appropriée des similarités textuelle et des citations, nous avons exploré et testé différentes formules dans notre approche. Nous avons expérimenté différentes techniques de pondération et de fusion pour combiner les deux types de similarités de manière adéquate.

Voici les formules de combinaison pour combiner les similarités de citation et de texte :

• **Formule de combinaison Min-Max :**

Calcule de la similarité minimale (Min) et maximale (Max) pour chaque paire de documents dans les matrices de similarité de citation et de texte. Utilisation de la formule suivante pour combiner les deux similarités :

$$Sim(di, dj) = \alpha * (\min txt\_sim(di, dj), cit\_sim(di, dj)) + (1 - \alpha) * \max(txt\_sim(di, dj), cit\_sim(di, dj)) \tag{3.6}$$

Avec :

- α : parametre entre 0 et .1
- Txt_sim(di,dj) : similarity a base du text entre di et dj.
- Cit_sim(di,dj) : similarité a base de citation entre di et dj.

• **Formule de combinaison par le Produit :**

Multiplication des deux matrices de similarité de citation et de texte.

Utilisation de la formule suivante pour normaliser la similarité combinée :

$$Sim(di, dj) = cit\_sim(di, dj) * txt\_sim(di, dj)^{\alpha} \tag{3.7}$$

• **Formule de combinaison par moyenne harmonique :**

Calcule la moyenne harmonique (Harmonic Mean) des similarités de citation et de texte pour chaque paire de documents. Utilisation la formule suivante pour combiner les similarités :

$$\text{Sim (di,dj)} = 2*(\text{Cit\_Sim(di,dj)}*\text{txt\_Sim(di,dj)})/(\text{Cit\_Sim(di,dj)}*\text{txt\_Sim(di,dj)}) \tag{3.8}$$

• **Combinaison non linéaire :**

Application une fonction non linéaire pour combiner les similarités de citation et de texte. Utiliser la fonction exponentielle ou la fonction sigmoïde pour pondérer les similarités :

$$Sim(di, dj) = exp(\alpha * Cit\_Sim(di, dj)*) * exp((1 - \alpha) * txt\_Sim(di, dj)) \tag{3.9}$$

• **Moyenne pondérée :**

Attribuez des poids à la similarité de citation et à la similarité de texte en fonction de leur importance relative. Multipliez la matrice de similarité de citation par son poids et la matrice de similarité de texte par son poids. Additionnez les deux matrices pondérées pour obtenir la matrice de similarité combinée.

$$Sim(di, dj) = \alpha * Cit\_Sim(di, dj) + (1 - \alpha) * txt\_Sim(di, dj) \tag{3.10}$$

• **Concaténation de matrices :**

Concaténez la matrice de similarité de citation et la matrice de similarité de texte horizontalement ou verticalement. Application d'une mesure de similarité (par exemple, la similarité cosinus) à la matrice concaténée pour obtenir la matrice de similarité combinée. Par exemple, si la matrice de similarité de citation est notée C et la matrice de similarité de texte est notée T, la matrice de similarité combinée peut être calculée comme suit :

$$Sim(di, dj) = np.concatenate(Cit\_Sim(di, dj), txt\_Sim(di, dj) \tag{3.11}$$

• **Normalisation et combinaison linéaire :**

Normalisez la matrice de similarité de citation et la matrice de similarité de texte pour obtenir des valeurs entre 0 et 1. Attribuez des poids aux matrices normalisées en fonction de leur importance relative. Effectuez une combinaison linéaire des matrices normalisées en utilisant les poids attribués. Par exemple, si la matrice de similarité de citation normalisée est notée NC et la matrice de similarité de texte normalisée est notée NT, et si vous attribuez un poids de 0,6 à la similarité de citation et un poids de 0,4 à la similarité de texte, la matrice de similarité combinée peut être calculée comme suit :

- cit_sim = (cit_sim- np.min(cit_sim)) / (np.max(cit_sim) - np.min(cit_sim))
- txt_Sim = (txt_Sim - np.min(txt_Sim)) / (np.max(txt_Sim) -np.min(txt_Sim))

$$Sim(di, dj) = (cit\_weight * cit\_sim(di, dj)) + (txt\_weight * txt\_Sim(di, dj)) \tag{3.12}$$

Voici un exemple d'affichage de la matrice de combinaison utilisant la moyenne pondérée :

3.4.2.5 Appliquer un algorithme de clustering

Après avoir effectué la représentation et la pondération des données ainsi que la combinaison des similarités textuelle et des citations, nous procédons à l'application d'un algorithme de clustering. Ce dernier est chargé de regrouper les documents en fonction de leurs similarités, en formant des clusters ou des groupes cohérents.

Parmi les différents algorithmes de clustering disponibles, nous avons choisi l'algorithme k-means pour notre étude. On a choisi cet algorithme pour plusieurs raisons :

• **Efficacité :** K-means est connu pour sa rapidité de calcul, ce qui le rend adapté à des ensembles de données volumineux. Étant donné que mon étude porte sur un grand nombre de documents scientifiques, il est essentiel d'avoir un algorithme capable de traiter rapidement ces données.

• **Simplicité :** K-means est relativement simple à comprendre et à implémenter. Il repose sur le principe de minimisation de la variance intra-cluster, ce qui en fait une méthode intuitive pour regrouper les données. En tant que chercheur, il est important d'utiliser un algorithme que je peux facilement comprendre et expliquer aux autres.

• **Clusters compacts :** K-means a tendance à produire des clusters compacts, ce qui signifie que les membres d'un même cluster sont généralement plus similaires les uns aux autres. Cela est important dans mon étude car je cherche à regrouper les documents qui partagent des similitudes textuelles et des liens de citation. Des clusters compacts nous aideront à identifier des groupes de documents étroitement liés sur le plan thématique.

• **Validation et interprétation :** K-means fournit des résultats facilement interprétables. Les centroïdes de chaque cluster peuvent être utilisés pour représenter les caractéristiques principales du groupe.

A. Nous rappelons ici le Fonctionnement de l'algorithme K-means :

1. Le nombre de clusters K est choisi à l'avance par l'utilisateur.

2. Les centres de cluster initiaux sont choisis aléatoirement parmi les points de données.

3. Chaque point de données est affecté au centre de cluster le plus proche (distance Euclidienne).

4. Les centres de cluster sont recalculés en utilisant la moyenne des points de données qui y sont affectés.

5. Les étapes 3 et 4 sont r´ep´et´ees jusqu'à ce que les centres de cluster ne bougent plus significativement ou que le nombre maximum d'itérations soit atteint.

L'algorithme converge généralement assez rapidement, mais il peut être sensible au choix initial des centres de cluster et peut aboutir 'a des optima locaux. Des variantes de l'algorithme, comme le K-means ++, ont été proposées pour résoudre ce probléme.

B. ici nous rappelons Le pseudo-algorithme de K-means :

Algorithm 5 Algorithme des K-Means

Entrée : $D = \{t_1, t_2, \ldots, t_n\}$ {Ensemble d'éléments}

K {Nombre de clusters souhaités}

Sortie : K {Ensemble de clusters}

Algorithme K-Means :

Assigner des valeurs initiales pour $m_1, m_2, \ldots, m_k$

Répéter

Assigner chaque élément t_i au cluster qui a la moyenne la plus proche

Calculer la nouvelle moyenne pour chaque cluster

Jusqu'à ce que le critère de convergence soit atteint

C. Comment trouver le meilleur k :

• Pour déterminer le meilleur K dans l'algorithme K-means, la méthode du coude (Elbow method) est utilisée. La méthode du coude (Elbow method) est une technique utilisée pour déterminer le nombre optimal de clusters dans un algorithme de clustering, tel que K-means. Elle tire son nom de la forme du graphique qui représente la relation entre le nombre de clusters et une métrique d'évaluation du clustering, généralement la somme des carrés des distances (SSE) ou l'inertie.

• L'idée principale de la méthode du coude est de choisir un nombre de clusters qui trouve un équilibre entre la réduction de la SSE (cohésion intra-cluster) et la complexité ajoutée par l'ajout de clusters supplémentaires. Plus précisément, à mesure que le nombre de clusters augmente, la SSE tend à diminuer car chaque cluster devient plus petit et plus concentré. Cependant, il existe un point où l'ajout d'un cluster supplémentaire n'apporte qu'une amélioration marginale de la SSE, car la majorité des données sont déjà bien regroupées. Ce point est souvent appelé le "coude" du graphique.

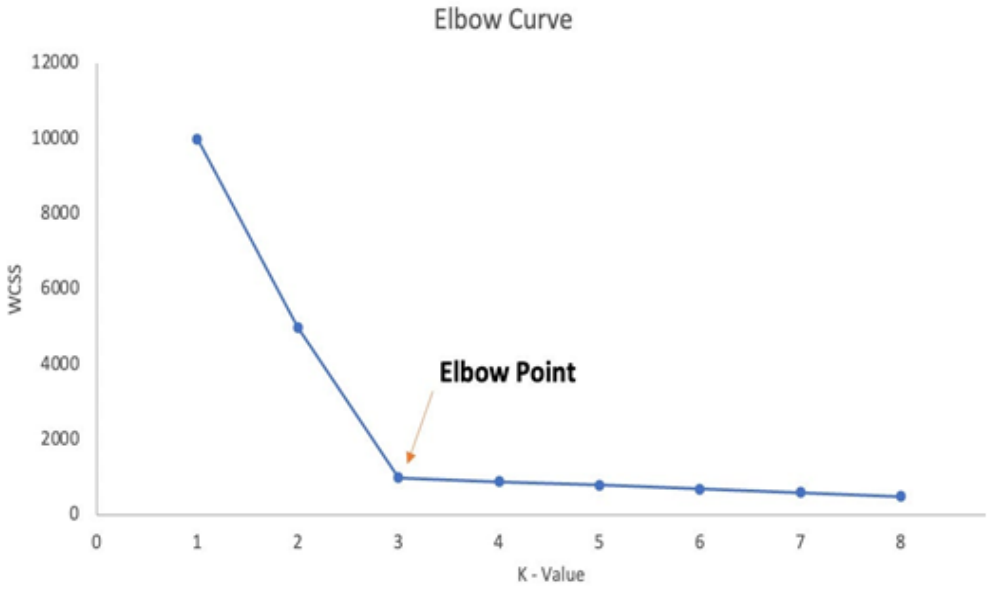

FIGURE 3.10 – Fonctionnement de la méthode du coude [10]

3.4.2.6 Détection d'un sujet pour chaque cluster

Après la phase de Clustering, une étape très importante reste à réaliser. Il s'agit de caractériser chaque Cluster afin de donner une indication sur le contenu (ou le thème) des articles qui le compose. Cette étape est plus que nécessaire afin de mieux organiser la collection des articles et aussi pour guider l'utilisateurs vers les articles de son intérêt.

Nous choisissons dans chaque Cluster, l'article le plus proche de son centroïde comme l'article le plus représentatif de ce cluster. Cette proximité est déterminée en calculant la mesure cosinus entre le vecteur du centroïde et les vecteurs des articles présents dans le cluster. Voici comment elle fonctionne :

• Tout d'abord, le centroïde du cluster est calculé en prenant la moyenne des vecteurs des articles qui composent ce cluster. Le centroïde représente le point central du cluster dans l'espace vectoriel.

• Ensuite, pour chaque article dans le cluster, on calcule la similarité cosinus entre le vecteur de l'article et le vecteur du centroïde.

• En comparant les valeurs de similarité cosinus obtenues pour chaque article, on identifie celui qui a la valeur la plus élevée. Cela signifie que cet article est le plus proche du centroïde en termes de similarité thématique et est donc considéré comme l'article le plus représentatif du cluster.

• Une fois que l'article le plus proche du centroïde est identifié, il est utilisé pour

caractériser le cluster en lui attribuant un titre qui reflète le sujet principal représenté par cet article.

Exemple :

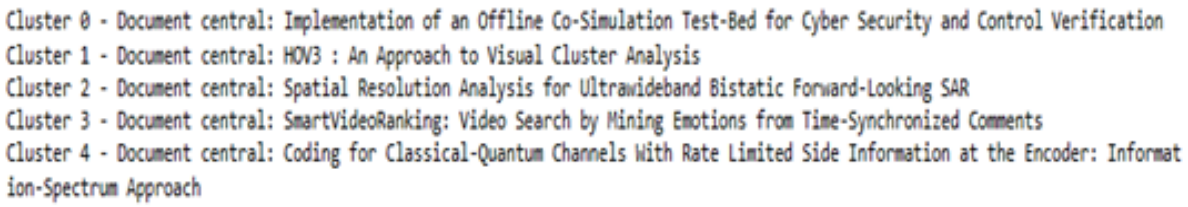

FIGURE 3.11 – Fonctionnement de la méthode du coude

3.4.2.7 Evaluation de notre clustering

Le score de silhouette est une mesure de la qualité d'un clustering. Il évalue à quel point chaque échantillon est similaire à son propre cluster par rapport aux autres clusters. Le score de silhouette varie de -1 à 1, où une valeur proche de 1 indique une bonne séparation des clusters, une valeur proche de 0 indique un chevauchement entre les clusters et une valeur proche de -1 indique une mauvaise séparation des clusters.

Le calcul du score de silhouette se fait en utilisant la distance entre les échantillons et la distance moyenne des échantillons dans leur cluster respectif. Le score de silhouette pour un échantillon est calculé comme suit :

1. Calculer la distance moyenne entre l'échantillon et tous les autres échantillons du même cluster. Cela donne la distance intra-cluster (a).

2. Calculer la distance moyenne entre l'échantillon et tous les échantillons d'un autre cluster. Cela donne la distance inter-cluster (b).

3. Calculer le score de silhouette pour l'échantillon en utilisant la formule :

$$\text{silhouette\_score} = \frac{b - a}{\max(a, b)}$$

.

Le score de silhouette global pour l'ensemble des échantillons est la moyenne des scores de silhouette individuels.

Un score de silhouette élevé indique que les échantillons sont bien regroupés et séparés les uns des autres, tandis qu'un score bas indique que les échantillons sont mal regroupés et qu'il y a un chevauchement entre les clusters. cette figure représente comment calculer score silhouette 3.12

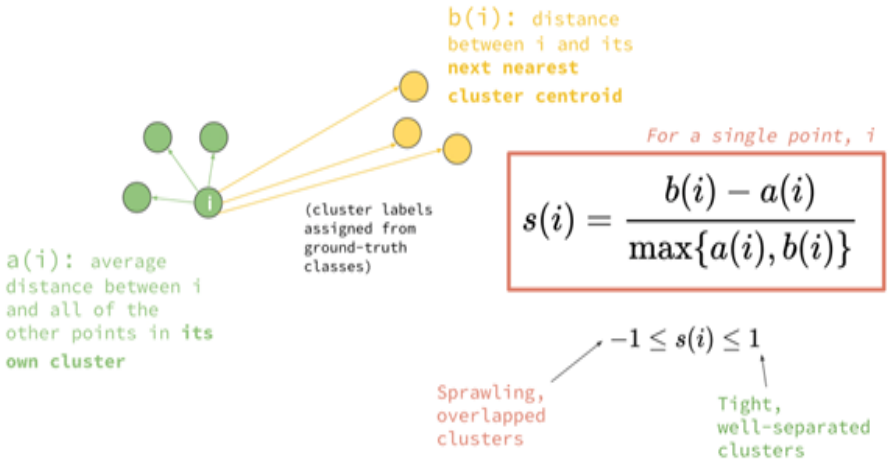

FIGURE 3.12 – Score sihlouette [7]

3.5 Conclusion

En conclusion de ce troisième chapitre, nous avons exposé notre approche détaillée pour effectuer le regroupement des articles scientifiques, nous avons d'abord examiné les travaux connexes dans le domaine, en soulignant les avancées récentes et les différentes approches existantes. Puis nous avons présenté les éléments de réflexion et les hypothèses sur lesquelles nous nous somme basé pour élaborer notre système. Nous avons surtout souligné la nécessité d'exploiter la richesse des documents scientifique en termes de structure, de métadonnées et de citations dans le processus de représentation et de clustering des articles. Nous avons aussi introduit une nouvelle formule de calcule de similarité à base de citation en prenant en compte la fréquence d'apparition des citations dans le document. Enfin nous avons essayés plusieurs formules de combinaison des similarités.

Dans le prochain chapitre, nous présenterons les expérimentations de notre solution sur un datasheet d'articles scientifique. Nous présenterons les résultats et nous les discuterons dans le détail.

Chapitre 4

Expérimentation et évaluation

4.1 Introduction

Dans ce chapitre se consacre à l'évaluation approfondie et à la présentation des résultats de notre méthode de regroupement de documents scientifiques. Après avoir exposé en détail notre approche de clustering, mettant en évidence la combinaison des deux similarités et l'utilisation de l'algorithme K-means, il est désormais essentiel de procéder à une évaluation approfondie afin de quantifier la qualité et l'efficacité de notre méthode.

A travers cette évaluation nous voulons confirmer (ou infirmer) nos éléments de réflexion de départ à savoir :

- L'intérêt de combiner la similarité à base des réseaux de citation dans le processus de clustering des documents scientifiques.

- L'intérêt de prise en charge des informations de structure (dans notre cas la rubrique titre) dans l'amélioration des résultats du Clustering.

- Enfin mesurer l'impact de de l'introduction de la fréquence des citations dans les formules de similarité à base de citation entre les documents.

4.2 Dataset utilisé

Le dataset utilisé dans notre expérimentation est un dataset destiné uniquement à des fins de recherche. Ce dataset comprend 3,079,007 articles et 25,166,994 citations. Chaque article est associé à un résumé, des auteurs, une année de publication, une source et un titre.

Ce jeu de données peut être utilisé pour des tâches de regroupement (clustering) avec des informations réseau et contextuelles, l'étude de l'influence dans le réseau de citations, l'identification des articles les plus influents, l'analyse de modélisation thématique, etc.

4.2.1 Exploration préliminaire des données

Nous avons utilisé pour ce travaille un dataset DBLP qui est adoptée par la communauté scientifique dans le but de rechercher et d'accéder à des publications de recherche dans le domaine de l'informatique. Ce dataset contient 08 attributs (variables), et chaque attribut contient des instances comme suit : 4.1

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 5000 entries, 2815 to 66845
Data columns (total 8 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   abstract    2855 non-null   object
 1   authors     5000 non-null   object
 2   n_citation  5000 non-null   int64
 3   references  3162 non-null   object
 4   title       5000 non-null   object
 5   venue       5000 non-null   object
 6   year        5000 non-null   int64
 7   id          5000 non-null   object
dtypes: int64(2), object(6)
memory usage: 351.6+ KB
```

FIGURE 4.1 – la structure d'un document scientifique

4.2.2 Description des attributs utilisées

Le tableau décrit les 08 attributs utilisés et précise leurs types leurs descriptions avec des exemples :

• 'abstract' : Le résumé du document scientifique, qui fournit un aperçu du contenu de l'article.

• 'authors' : Les auteurs du document, répertoriant les individus ou les groupes de recherche responsables de la recherche.

• 'n_citation' : Le nombre de fois où le document a été cité par d'autres articles, indiquant son impact ou son influence dans la communauté scientifique.

• 'references' : Les références citées dans le document, répertoriant d'autres articles ou sources qui ont été référencés dans la recherche.

• 'title' : Le titre du document scientifique, qui donne une brève description du sujet de recherche.

• 'venue' : Le lieu de publication ou la revue dans laquelle le document a été publié.

• 'year' : L'année de publication du document.

• 'id' : Un identifiant unique pour chaque document, qui peut être utilisé pour les références ou les liens.

• 'num_references' : Le nombre total de références citées dans le document.

Field Name	Field Type	Description	Example
id	string	paper ID	013ea675-bb58-42f8-a423-f5534546b2b1
title	string	paper title	Prediction of consensus binding mode geometries for related chemical series of positive allosteric modulators of adenosineand muscarinic acetylcholine receptors
authors	list of strings	paper authors	["Leon A. Sakkal", "Kyle Z. Rajkowski", "Roger S. Armen"]
venue	string	paper venue	Journal of Computational Chemistry
year	int	published year	2017
n_citation	int	citation number	0
references	list of strings	citing papers' ID	["4f4f200c-0764-4fef-9718-b8bccf303dba", "aa699fbf-fabe-40e4-bd68-46eaf333f7b1"]
abtract	string	abstract	This paper studies ...

TABLE 4.1 – Les données de dataset.

4.3 Evaluation

Nous avons réalisé plusieurs expérimentations afin d'obtenir une évaluation précise de notre approche. Pour ce faire, nous avons évalué plusieurs variantes ou stratégies différentes. Nous avons défini ces différentes variantes en fonction des paramètres suivants :

- L'utilisation des rubriques titre et résumé ensemble dans la représentation des documents scientifiques.

- Pondération préférentielle des mots apparus dans la rubrique titre.

- L'utilisation de différentes formules de combinaison entre les deux similarités textuelles et citation.

Dans notre approche, le clustering des documents scientifiques est réalisé en utilisant différentes approches qui prennent en compte à la fois la similarité textuelle et la similarité des citations. Pour évaluer la similarité textuelle, nous avons exploré différentes méthodes de calcul basées sur le vocabulaire utilisé dans les titres et les résumés des articles.

En ce qui concerne la similarité des citations, nous avons testé plusieurs formules de combinaison qui intègrent les informations sur les références et les sources citées dans les articles.

Les résultats des différentes variantes expérimentales, évaluées en utilisant le coefficient de silhouette comme mesure de performance, calculé à l'aide de la bibliothèque Scikit-learn de Python avec la fonction " score_silhouette".

4.3.1 Variantes selon la similarité textuelle uniquement

Dans la première évaluation, nous avons considéré uniquement la similarité textuelle des documents pour regrouper les articles. Nous n'avons pas pris en compte la similarité à base de citations. Cela signifie que nous avons mesuré la similarité des contenus textuels des articles en utilisant la mesure de similarité « cosinus ». nous avons ici évaluer deux variantes selon l'importance accodrée au termes qui figure dans la rebrique « titre » :

A- Vatiante_Titre_résumé :

Nous avons utilisé le titre et le résumé de chaque article pour représenter et calculer la similarité entre les documents. En utilisant ces deux composantes textuelles, nous avons pu mesurer à quel point les articles étaient similaires les uns aux autres. Le titre donne un aperçu concis du sujet de l'article, tandis que le résumé fournit une description plus détaillée du contenu. En combinant ces deux éléments, nous avons pu capturer les informations clés et évaluer la similitude entre les articles en fonction de leur contenu textuel.

B- Variante_titre*n_résumé :

Dans notre approche, nous avons cherché à accorder une importance supplémentaire

aux termes qui figure dans le titre des articles lors du calcul de la similarité. Nous avons réalisé cela en utilisant une pondération plus élevée pour le titre par rapport au résumé. Le titre d'un article est souvent plus concis et représentatif de son contenu global, fournissant des informations clés sur le sujet traité. En attribuant un poids plus élevé au titre, nous avons cherché à mettre en avant les similitudes entre les titres des articles, ce qui nous permet d'identifier plus facilement les documents qui partagent des thématiques communes. Nous estimons que cette pondération a contribué à améliorer la précision de notre méthode de similarité textuelle en mettant en évidence les articles qui sont étroitement liés en termes de contenu, mais qui peuvent présenter des variations dans leurs résumés.

4.3.2 Variante selon similarité combinée texte et citation

Dans ce deuxième groupe de variantes évaluées, nous avons combiné les similarités à base de contenu textuel et à base de citation. Nous avons exploré différentes formules de combinaison pour prendre en compte à la fois la similarité textuelle et la similarité des citations lors du clustering des documents scientifiques. Ces formules nous ont permis de fusionner les mesures de similarité provenant des deux aspects afin d'obtenir une mesure globale de similarité entre les documents. Nous avons expérimenté différentes approches, telles que l'attribution de poids plus importants à la similarité des citations par rapport à la similarité textuelle, ou vice versa. L'objectif de ces expérimentations était de trouver la meilleure formule de combinaison qui capture efficacement les deux aspects de similarité et qui permet de regrouper les documents de manière cohérente et pertinente.

Variante_titre*n_résumé_citation : Dans cette variante nous avons combiné la similarité à base du texte selon la variante Variante_titre*n_résumé avec une similarité à base de citation selon plusieurs formules de combinaison (min-max, moyenne pondérée, combinaison linéaire...).

Nous donnerons dans ce qui suit les différents résultats des expérimentations :

4.3.3 Résultats des expérimentations

Tout d'abord, pour les variantes à base de similarité textuelle seulement :

	Score silhouette
Titre + résumé	0.55
Variante_titre*2_résumé	0.64

TABLE 4.2 – les résultats à base de similarité textuelle

Nous remarquons ici que le fait de doubler seulement la fréquence des termes qui figure dans la rubrique titre permet d'améliorer les résultats du Clustering.

Pour les résultats des équations pour les variantes à base de combinaison similarité textuelle et citations :

Selon la formule de combinaison min max :

$$Sim(di, dj) = \alpha * (\min txt\_sim(di, dj), cit\_sim(di, dj)) + (1 - \alpha) * \max(txt\_sim(di, dj), cit\_sim(di, dj)) \tag{4.1}$$

	Indice de silhouette		
Variante	a =0.2	a=0.5	a = 0.8
Variante _titre*2_résumé_citation	0.57	0.59	0.59

TABLE 4.3 – Selon la formule de combinaison min max

Selon la formule de moyenne pondérée :

$$Sim(di, dj) = \alpha * Cit\_Sim(di, dj) + (1 - \alpha) * txt\_Sim(di, dj) \tag{4.2}$$

	Indice de silhouette		
Variante	a =0.2	a=0.5	a = 0.8
Variante _titre*2_résumé_citation	0.54	0.59	0.61

TABLE 4.4 – Selon la formule de moyenne pondérée

Selon la formule de combinaison non linéaire :

	Indice de silhouette		
Variante	a =0.2	a=0.5	a = 0.8
Variante _titre*2_résumé_citation	0.58	0.60	0.70

TABLE 4.5 – Selon la formule de combinaison non linéaire

$$Sim(di, dj) = exp(\alpha * Cit\_Sim(di, dj)*) * exp((1 - \alpha) * txt\_Sim(di, dj)) \tag{4.3}$$

Selon la formule de normalisation et combinaison linéaire :

$$Sim(di, dj) = \alpha * Cit\_Sim(di, dj) + (1 - \alpha) * txt\_Sim(di, dj) \tag{4.4}$$

avec :

- cit_sim = (cit_sim- np.min(cit_sim)) / (np.max(cit_sim) - np.min(cit_sim))

- txt_Sim = (txt_Sim - np.min(txt_Sim)) / (np.max(txt_Sim) -np.min(txt_Sim))

	Indice de silhouette		
Variante	a =0.2	a=0.5	a = 0.8
Variante _titre*2_résumé_citation	0.58	0.60	0.71

TABLE 4.6 – Selon la formule de normalisation et combinaison linéaire

Nous remarquons que la variantes avec combinaison des similarité textuelles et de citation donne de meilleurs résultats que l'approche à base de similarité textuelle seulement. Meme si le poids de la similarité textuelle reste le plus important (0.8) dans les configurations les plus performantes (indice de silhouette >70%).

4.4 Discussion des résultats

Dans notre étude, nous avons utilisé une approche de clustering basée sur la similarité textuelle, en prenant en compte à la fois le titre et le résumé des articles scientifiques. Les résultats obtenus à partir de cette approche initiale de clustering ont montré une certaine capacité à regrouper les articles selon des thèmes ou des sujets communs. Cependant, nous avons constaté que l'ajout de la similarité des citations a permis d'améliorer la qualité des regroupements obtenus.

Lorsque nous avons intégré la similarité des citations dans le processus de clustering, nous avons observé une augmentation significative de la cohérence des clusters formés. Cela s'explique par le fait que les articles partageant des références communes ont tendance à être regroupés ensemble, ce qui renforce la cohésion des clusters en termes de contenu scientifique. Et D'après les résultats obtenus, il est clair que l'approche de combinaison hybride se distingue par sa précision élevée en comparaison avec d'autres approches de clustering qui se basent uniquement sur la similarité textuelle.

4.5 Environnement de test

4.5.1 Langage de programmation

Pour la réalisation de notre solution, nous avons mis en œuvre notre approche en utilisant le langage de programmation Python.

Python est un langage de programmation interprété, de haut niveau et polyvalent. Il a été créé par Guido Van Rossum et sa première version a été publiée en 1991. Python se distingue par sa syntaxe claire et lisible, ce qui facilite l'écriture et la compréhension du code.[48]

4.5.2 Bibliothèques utilisés

Scikit-Learn

Scikit-learn est une bibliothèque Python qui fournit une interface standard pour la mise en œuvre d'algorithmes d'apprentissage automatique, crée par David Cournapeau, elle construite en utilisant les libraires NumPy, SciPy, et matplotlib de python. Scikit-Learn propose une large gamme d'algorithmes d'apprentissage automatique, tels que la régression linéaire, la régression logistique, les machines à vecteurs de support (SVM), les arbres de décision, les forêts aléatoires, les k-plus proches voisins (KNN), le clustering (regroupement), et bien d'autres.

Scikit-Learn propose également des fonctionnalités pour extraire la structure de données complexes telles que les bases de données, les textes et les images, afin de les classer à l'aide de techniques telles que la méthode de pondération TF-IDF.[49]

Pandas

Pandas est une bibliothèque Python riche en fonctionnalités, conçue pour la manipulation et l'analyse de données structurées. Elle offre des structures de données performantes et des outils intuitifs pour effectuer des manipulations courantes sur les ensembles de données. Pandas est largement utilisé dans des domaines tels que les statistiques, la finance et les sciences sociales. Son objectif est de devenir la couche fondamentale pour les calculs statistiques en Python, complétant ainsi l'écosystème existant des outils scientifiques. Il ouvre de nouvelles perspectives de développement et de croissance pour les applications d'analyse de données en Python.[50]

NLTK

NLTK est une bibliothèque Python complète pour le traitement et l'analyse du langage naturel. Avec des fonctionnalités telles que la tokenisation, la lemmatisation, le stemming, l'analyse syntaxique et la classification de texte, NLTK est largement utilisé dans l'industrie et la recherche pour ses capacités étendues. Il est particulièrement apprécié pour son utilité dans le prétraitement des textes et son large éventail d'outils disponibles.[51]

Matplotlib

Matplotlib est une bibliothèque Python utilisée pour créer des graphiques et des visualisations de données. Elle offre une grande flexibilité et permet de créer une variété de graphiques, tels que des diagrammes en barres, des courbes et des histogrammes. Matplotlib est souvent utilisé avec d'autres bibliothèques comme NumPy et Pandas pour l'analyse de données et la création de visualisations. C'est un outil essentiel dans les domaines de la science des données, de la recherche et de l'apprentissage automatique où la représentation visuelle des données est importante.[52]

Spacy

Est une bibliothèque réalisée en Python et Cython en 2015, puissante et spécialisée dans le traitement du langage naturel. Elle propose des fonctionnalités avancées pour l'analyse de texte, telles que la tokenisation précise, la lemmatisation des mots, l'analyse syntaxique détaillée et la reconnaissance des entités nommées. Spacy se distingue par sa performance élevée et sa facilité d'utilisation.s Cette bibliothèque est largement utilisée dans la recherche, l'analyse de données, l'apprentissage automatique et d'autres domaines où une compréhension précise et approfondie du langage naturel est essentielle.[53]

4.5.3 plateforme et environnement de test

Anaconda

Anaconda est une distribution logicielle populaire et gratuite pour les langages de programmation Python et R. Elle est largement utilisée dans le domaine de la science des données et du calcul numérique. Anaconda fournit un environnement complet et prêt à l'emploi, regroupant de nombreuses bibliothèques populaires telles que NumPy, Pandas, Matplotlib, SciPy, scikit-learn et bien d'autres, qui sont couramment utilisées dans l'analyse de données et l'apprentissage automatique. De plus, il inclut également des outils comme Jupyter Notebook, qui permet aux utilisateurs de créer et d'exécuter du code Python de manière interactive, facilitant ainsi la collaboration et le partage de résultats.[54]

Kaggel

Kaggle, une filiale de Google, est une plateforme communautaire destinée aux scientifiques et aux développeurs de données. Elle rassemble une communauté dynamique de plus d'un million d'utilisateurs enregistrés provenant de 194 pays avec plus de 536 000 membres actifs, offrant ainsi un espace propice aux discussions sur l'apprentissage automatique, les développements modernes et l'exploration d'ensembles de données. Cette plateforme permet aux membres d'échanger leurs connaissances, de collaborer sur des modèles de développement, et de tisser des liens professionnels dans un environnement mondial stimulant.[55]

4.6 Conclusion

En conclusion, ce chapitre a été consacré à la réalisation de notre approche, nous avons configuré et testé plusieurs variantes expérimentales afin de trouver la meilleure solution pour notre cas d'étude. L'évaluation de chaque variante a été réalisée en utilisant la mesure d'évaluation silhouette. Finalement, nous avons procédé à une description détaillée de l'environnement de développement, des outils utilisés et du langage de programmation choisi pour la mise en œuvre de notre solution.

Conclusion générale

Le travail présenté dans ce mémoire représente notre projet de fin d'étude de Master. Il était question dans notre projet de développer un système de Clustering de documents scientifiques pour permettre le regroupement automatique d'une collection d'articles scientifique selon les thèmes et sujets abordés dans ces articles.

L'intérêt d'un tel système pour la communauté universitaires est évident. En effet devant la grande masse de publications scientifiques en continuel croissance, il devient difficile pour un utilisateur d'explorer facilement les collections d'articles scientifiques et de trouver rapidement les articles qui correspond à son intérêt. Les techniques de machine Learning, en particulier le Clustering s'impose comme une solution à cette problématique. Le Clustering permet de regrouper automatiquement ensemble des articles scientifiques selon leurs similarités calculé sur la base de la représentation de chaque article.

Il se trouve que pour les documents scientifiques, plusieurs sources d'évidences et de caractéristiques peuvent être exploitées en plus de leurs contenus scientifiques. Parmi ces composantes les plus importante on cite les liens de citations entre les articles. L'analyse des réseaux de citations nous permet aussi de calculer la similarité entre articles scientifiques sur la base de leurs liens de référence.

Nous avons donc développé un système de Clustering de documents scientifiques en se basant sur une approche hybride qui combine la similarité à base de contenu textuel et la similarité à base de liens de citation. Nous avons introduit le facteur de fréquence des citations dans le calcule de la similarité à base de citation.

En combinant ces deux sources d'informations complémentaires, nous avons cherché à améliorer la pertinence et la cohérence des regroupements obtenus, ainsi qu'à faciliter l'accès aux connaissances scientifiques.

De plus, nous avons évalué différentes variantes de notre approche en considérant les différentes composantes des articles scientifiques, telles que le titre, le résumé et les références. Cette évaluation approfondie nous a permis de mieux comprendre l'impact de chaque composante sur la qualité des regroupements obtenus.

Bibliographie

[1] https://www.facebook.com/GuideEnseignantEtChercheur/photos/a.853816604713554/1983381341757069/?type=3. Consulté le 11/03/2023.

[2] S. Zhang, Y. Xu, and W. Zhang. "clustering scientific document based on an extended citation model". *IEEE Access*, 7 :57037–57046, 2019.

[3] https://www.sap.com/products/artificial-intelligence/what-is-machine-learning.html. Consulté le 12/05/2023.

[4] Guillaume Saint-Cirgue. "Comment fonctionne le Machine Learning ?". 2020.

[5] Hausmane Issarane. "Le Clustering : Définition et Top 5 Algorithmes - BrightCape", 2019.

[6] https://slideplayer.com/slide/4829376/. Consulté le 05/04/2023.

[7] https://developers.google.com/machine-learning/clustering/clustering-algorithms?hl=fr. Consulté le 18/04/2023.

[8] Emre Yesilyurt. What is the Difference Between Hierarchical and Partitional Clustering. 2022.

[9] https://www.javatpoint.com/clustering-in-machine-learning. Consulté le 19/03/2023.

[10] Anmol Tomar. "Stop Using Elbow Method in K-means Clustering, Instead, Use this !". 2023.

[11] https://coop-ist.cirad.fr/rediger/article-scientifique/. Consulté le 12/05/2023.

[12] https://www.scribbr.fr/category/article-scientifique/. Consulté le 11/03/2023.

[13] J. Brooks. "the basic characteristics of scientific english", 2022.

[14] `https://www.anl.gov/education/features-of-good-scientific-writing/`. Consulté le 11/03/2023.

[15] `https://www.cwauthors.com/article/Different-types-of-scientific-papers/`. Consulté le 11/03/2023.

[16] La structure d'un article scientifique. umc.edu.dz, 2015. Consulté le 25/02/2023.

[17] `https://www.bibl.ulaval.ca/services/citation-de-sources`. consulté le 09/05/2023.

[18] Jack Caulfield. "Citation Styles Guide — Examples for All Major Styles". *Scribbr*, 2022.

[19] Justine Debret. "Aperçu des styles de citation". *Scribbr*, 2020.

[20] Filippo Radicchi, Santo Fortunato, and Alessandro Vespignani. *"Citation Networks"*. Springer Nature, 2011.

[21] Shuai Zhang, Yangbing Xu, and Wenyu Zhang. "Clustering Scientific Document Based on an Extended Citation Model". *IEEE Access*, pages 57037–57046, 2019.

[22] MIT Sloan. "machine learning, explained". 2021.

[23] `https://datascientest.com/machine-learning-tout-savoir`.

[24] Samira Toucherifte. *"Étude comparative en classification non-supervisée"*. Thèse de doctorat, 2011.

[25] I. Belcic. *"Qu'est-ce qu'un spam : le guide essentiel de détection et de prévention des spams"*. 2023.

[26] Youssif Al-Nashif, Aarthi Arun Kumar, Salim Hariri, Yi Luo, Ferenc Szidarovsky, and Guangzhi Qu. "multi-level intrusion detection system (ml-ids)". pages 131–140. IEEE, 2008.

[27] Jonathan G. Richens, Ciarán M. Lee, and Saurabh Johri. "improving the accuracy of medical diagnosis with causal machine learning". 2020. Corrected publication 2021.

[28] Ayoub Abraich, Hoang Dung Nguyen, and Mohamed Tounsi. "détection de fraude ieee-cis". 2020.

[29] `https://dataanalyticspost.com/Lexique/clustering/`. Consulté le 18/03/2023.

[30] `https://stacklima.com/exploration-de-donnees-analyse-de-cluster/`. Consulté le 10/04/2023.

[31] Nesrine Masmoudi. *"Modèle bio-inspiré pour le clustering de graphes : application à la fouille de données et à la distribution de simulations"*. Thèse de doctorat, Université de Normandie ; Université de Sfax (Tunisie), 2017.

[32] Abderraouf Boukhatem, Alexandre Duhamel, and David-Alexandre Eklo. "Étude de méthodes de clustering pour la segmentation d'images faciales". Université Paris-Dauphine, 2017.

[33] Saket Anand, Sushil Mittal, Oncel Tuzel, and Peter Meer. "semi-supervised kernel mean shift clustering". *IEEE transactions on pattern analysis and machine intelligence*, 36 :1201–1215, 2013.

[34] Maha Ghribi, Pascal Cuxac, Jean-Charles Lamirel, and Alain Lelu. "mesures de qualité de clustering de documents : Prise en compte de la distribution des mots clés". 2010.

[35] El-khadir Lamrani, El Habib Benlahmar, Hammad Ballaoui, Abdelaziz Marzak, and Kamal El Guemmat. "méthodes de clustering de documents textes arabes : étude comparative". *Université Hassan II - Mohammedia, Faculté des Sciences, Maroc*, 2014.

[36] Pankaj Jajoo. "document clustering". *IIT Kharagpur, Thesis*, 2008.

[37] Ricco Rakotomalala. "construction de la matrice documents termes". Cours université Lyon2, 2016.

[38] Raheel Saeed. *"L'Apprentissage Artificiel pour la Fouille de Données Multilingues"*. PhD thesis, 2010.

[39] Oumaima Alaoui Ismaili. *"Clustering prédictif : Décrire et prédire simultanément"*. PhD thesis, Université Paris Saclay (COmUE), 2016.

[40] Julien Ah-Pine. "une famille d'indices de similarité généralisant la mesure de cosinus". 2010.

[41] Ch Suquet. "distances euclidiennes sur les mesures signées et application à des théorèmes de berry-esséen". *Bulletin of the Belgian Mathematical Society-Simon Stevin*, pages 161–181, 1995.

[42] Christine Largeron, Bernard Kaddour, and Maria P Fernandez. "softjaccard : une mesure de similarité entre ensembles de chaînes de caractères pour l'unification d'entités nommées". 2009.

[43] Stéphane Lallich and Philippe Lenca. "indices de qualité en clustering". In *Journée thématique : clustering et co-clustering*. Société française de classification, 2015.

[44] Pedram Vahdani Amoli and Omid Sojoodi Sh. "scientific documents clustering based on text summarization". Technical report, Faculty of Electrical, Computer and IT Engineering, Qazvin Islamic Azad University, 2015.

[45] Rupesh Kumar Mishra, Kanika Saini, and Sakshi Bagri. "text document clustering on the basis of inter passage approach by using k-means". 2015.

[46] Thanh-Trung Van and Michel Beigbeder. "co-citations sur le web : Recherche de similarité entre les articles scientifiques". 2007.

[47] D. Yu, W. Wang, S. Zhang, W. Zhang, and R. Liu. "hybrid self-optimized clustering model based on citation links and textual features to detect research topics". *PLOS ONE*, 2017.

[48] KR Srinath. "python–the fastest growing programming language". *International Research Journal of Engineering and Technology*, pages 354–357, 2017.

[49] Ekaba Bisong and Ekaba Bisong. "introduction to scikit-learn". *Building Machine Learning and Deep Learning Models on Google Cloud Platform : A Comprehensive Guide for Beginners*, pages 215–229, 2019.

[50] Wes McKinney et al. "pandas : a foundational python library for data analysis and statistics". *Python for high performance and scientific computing*, pages 1–9, 2011.

[51] Meng Wang and Fanghui Hu. "the application of nltk library for python natural language processing in corpus research". *Theory and Practice in Language Studies*, pages 1041–1049, 2021.

[52] Igor Stančin and Alan Jović. "an overview and comparison of free python libraries for data mining and big data analysis". pages 977–982. IEEE, 2019.

[53] Eleni Partalidou, Eleftherios Spyromitros-Xioufis, Stavros Doropoulos, Stavros Vologiannidis, and Konstantinos Diamantaras. "design and implementation of an open source greek pos tagger and entity recognizer using spacy". pages 337–341, 2019.

[54] Akhil Kadiyala and Ashok Kumar. "applications of python to evaluate environmental data science problems". *Environmental Progress & Sustainable Energy*, pages 1580–1586, 2017.

[55] Casper Solheim Bojer and Jens Peder Meldgaard. "kaggle forecasting competitions : An overlooked learning opportunity". *International Journal of Forecasting*, pages 587–603, 2021.