


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 : *Ingénierie des systèmes d'information et logiciel*

Thème

Aggregated search engine for scientific publications

Encadré par

— M. BAL Kamel.

Réalisé par

— Mlle Debbi Sara

— Mlle Hellal Riane

2021/2022

Remerciements

Tout d'abord, Nous remercions Dieu qui nous a donné la force et la volonté pour la réalisation de ce travail.

Nous adressons nos profonds remerciements à nos chers parents pour leur Patience, Amour, Soutien et Encouragement.

Nous voudrions spécialement remercier, notre encadreur de mémoire **M.Kamal Bal** pour sa patience, sa disponibilité et surtout ses judicieux conseils, qui nous ont été précieux afin de mener notre travail à bon port.

Nous remercions département informatique, enseignants et administratifs qui nos aidé tout au long de mon périple à l'Université.

Nos vifs remerciements vont également aux membres du Jury qui nos 'ont fait l'honneur d'examiner et d'évaluer ce travail.

Enfin, nous tenons également à remercier toutes les personnes qui ont participé de près et loin à la réalisation de ce travail.

Dédicaces

Je dédie ce mémoire à mes parents pour leur amour inestimable, leur confiance, leur soutien, leurs sacrifices et toutes les valeurs qu'ils ont su m'enseigner.

Je dédie ce mémoire à mes chères sœurs ” **HANANE et TIZIRI** ” et mes frères ” **SOFIANE et FARID**” et ma famille ”**DEBBI** ” a ma chère copine asma et mon binome Riane

Je dédie ce mémoire à tous mes chers amis sans a toutes les personnes qui ont participé de manière directe ou indirecte à la concrétisation de ce travail.

DEBBI Sarah.

Dédicaces

. Je dédie ce mémoire à mes parents pour leur amour inestimable, leur confiance, leur soutien, leurs sacrifices et toutes les valeurs qu'ils ont su m'enseigner.

Je dédie ce mémoire à mon chère frère" **Islam**" et ma famille "**HELLAL** "et mon binome Sarah

Je dédie ce mémoire à tous mes chers amis sans a toutes les personnes qui ont participé de manière directe ou indirecte à la concrétisation de ce travail.

HELLAL Riane.

ملخص

البحث عن المقالات والمنشورات العلمية هو نشاط يومي أساسي في المجتمع العلمي. سواء كانوا معلمين أو باحثين أو طلابًا أو مبادرين لمشاريع مبتكرة ، فإن هذا النشاط يشغل جزءًا كبيرًا من وقتهم وعملهم. إن الكم الهائل من المعلومات العلمية التي يتم إنتاجها سنويًا ، وتنوع المصادر وتعددها ومحركات البحث المتخصصة يعني أن هذا النشاط للبحث عن المعلومات العلمية ذات الصلة أصبح معقدًا بشكل متزايد ويتطلب الكثير من الجهد والوقت.

استرجاع المعلومات المجمعة هو نموذج بحث يهدف إلى تجميع المعلومات من مصادر متعددة في نفس مساحة البحث. يتجاوز هذا النموذج استعادة المعلومات ذات الصلة ، ولكنه يطبق تقنيات التجميع من أجل إنتاج مساحة نتائج مجمعة يمكن للباحث عن المعلومات استغلالها بسهولة.

لقد حاولنا في هذا العمل المتواضع تطبيق هذا النموذج في مجال استرجاع المعلومات العلمية. اخترنا التجميع كأسلوب تجميع. لهذا ، قمنا باستغلال مصطلحات الأقسام : العنوان والملخص والكلمات الرئيسية لتمثيل المقالات نظرًا لأهمية هذه الأقسام في توضيح الموضوعات ومحتوى المقال العلمي.

لقد طبقنا خوارزمية كمان ثلسترنگ على مساحة بيانات تحتوي فقط على المصطلحات التي اخترناها. أخيرًا قمنا بتقييم عملنا باستخدام معامل الصورة الظلية لقياس جودة التجميع.

. **الكلمات المفتاحية** : بحث في المعلومات العلمية ، مقال علمي ، علاقات دولية مجمعة ، نظام بحث ، تجميع ، تكتل.

Résumé

La recherche d'articles et de publications scientifiques est une activité quotidienne essentielle au sein de la communauté scientifique. Qu'ils soient enseignants, chercheurs, étudiants ou porteurs de projets innovants, cette activité occupe une grande partie de leur temps et de leur travail. La grande quantité d'informations scientifiques produite annuellement, la diversité et la multitude des sources et moteurs de recherche spécialisés fond que cette activité de recherche d'information scientifique pertinente est devenu de plus en plus complexe et nécessite énormément d'effort et de temps.

La recherche d'information agrégée est un paradigme de recherche qui vise à assembler dans la même espace de recherche des informations issues de plusieurs sources. Ce paradigme va au delà de la restitution des informations pertinentes, mais appliques des techniques d'agrégation afin de produire un espace de résultats agrégé facilement exploitable par le chercheur d'information.

Nous avons dans ce modeste travail essayé d'appliquer ce paradigme dans le domaine de la recherche d'information scientifique. Nous avons opté pour de Clustering comme technique d'agrégation. Pour cela, nous avons exploité les termes des sections : titre, résumé et mots-clés pour la représentation des articles vu l'importance de ces sections dans l'indication du sujets et du contenu de l'article scientifique.

Nous avons appliqué l'algorithme de Clustering K-means sur un espace de données qui contient uniquement les termes qu'on a choisi. Finalement nous avons évalué notre travail en utilisant le coefficient de silhouette pour mesurer la qualité du Clustering.

Mots-clés : Recherche d'information scientifique, Article scientifique, RI agrégée, Système de recherche, agrégation, Clustering.

Abstract

Searching for scientific articles and publications is an essential daily activity within the scientific community. Whether they are teachers, researchers, students or initiators of innovative projects, this activity occupies a large part of their time and work.The large amount of scientific information produced annually, the diversity and multitude of sources and specialized search engines means that this activity of searching for relevant scientific information has become increasingly complex and requires a lot of effort and time.

Aggregated information retrieval is a search paradigm that aims to assemble information from multiple sources in the same search space.This paradigm goes beyond the restitution of relevant information, but applies aggregation techniques in order to produce an aggregated results space easily exploitable by the information seeker.

We have in this modest work tried to apply this paradigm in the field of scientific information retrieval. We opted for Clustering as an aggregation technique. For this, we have exploited the terms of the sections : title, abstract and keywords for the representation of the articles given the importance of these sections in indicating the subjects and the content of the scientific article.

We applied the K-means Clustering algorithm on a data space that contains only the terms that we have chosen. Finally we evaluated our work using the silhouette coefficient to measure the quality of the Clustering.

Keywords : Research of scientific information, Scientific article, Aggregated IR, Search system, aggregation, Clustering.

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

RI	Recherche d'information.
JSON	JavaScript Object Notation.
TF	Term Frequency.
IDF	Inverse Term Frequency.
SVD	Singular Value Decomposition.
BSD	Berkeley Software Distribution.
NLTK	Natural Language Toolkit.
GPU	Graphics Processing Unit.

Introduction générale

- ***Contexte et problématique***

Depuis son avènement, l'Internet a accéléré le développement d'applications dans tous les domaines. Les réseaux informatiques, qui constituent une infrastructure d'échange de plus en plus performante, ont permis une forte dissémination d'informations, de services et d'applications sur des serveurs distants à des volumes impressionnants.
La recherche d'articles et de publications scientifique est une activité quotidienne primordiale pour toute la communauté scientifique. Qu'ils s'agissent d'enseignants, de chercheurs, d'étudiants ou de porteurs de projet innovants, cette activité occupe une grande partie de leurs temps et de leur travail. La grande masse d'information scientifique produite chaque année et la multitude des sources et de moteurs de recherche spécialisés offrent à la communauté scientifique une grande commodité et les aident dans leur activité de recherche.
Cependant, cette grande masse d'informations recueillies par les moteurs de recherche scientifiques posent également de grands problèmes aux chercheurs. En effet, un chercheur qui soumet sa requête à ces moteurs de recherche reçoit un très grand nombre de résultats (des milliers d'articles des fois). Il devient très difficile pour lui de déceler dans tout cet espace de résultats l'information pertinente qui l'intéresse. Il est souvent obligé de lire les contenus de plusieurs articles et des fois de reformuler sa requête (la préciser) pour pouvoir tomber sur le bon article qu'il souhaite. Il dépense énormément de temps et d'énergie pour trouver l'information qu'il cherche. Cette difficulté est due au fait que ces moteurs de recherche de limitent seulement à présenter un long listing de liens vers les articles susceptibles d'être pertinents pour l'utilisateur. L'espace de résultats ainsi conçu

ne donne pas des éléments au chercheur pour le guider vers l'information souhaité ni pour mieux explorer cet espace de résultats. L'existence de plusieurs collections et de moteurs de recherche scientifiques obligent aussi le chercheur à les parcourir afin de comprendre le domaine qui l'intéresse.
La recherche d'information agrégée est un nouveau paradigme de la RI qui permet d'interroger plusieurs sources d'information et de présenter à l'utilisateur un espace de réponse bien organisée afin de l'aider à bien explorer cet espace de résultats. La RI agrégée ne se limite pas à lister seulement les résultats de recherche issues de plusieurs sources mais s'intéresse aussi à l'agrégation des résultats de recherche en cherchant les liens qui peuvent exister entre les différents résultats dans le but de les organiser et de permettre une meilleure visualisation de l'espace de résultats. Nous envisageons d'exploiter ce paradigme pour répondre aux problèmes cités plus haut. Nous projetons de développer un système de recherche d'information agrégée pour la recherche d'articles scientifique.

Objectifs

Nous envisageons de développer un système de recherche d'information agrégée pour la recherche d'articles scientifique qui permettra de :
-Aider le chercheur dans son activité de recherche.
- Exploiter la spécificité les articles scientifique pour mieux organiser l'espace des résultats.
-Exploiter l'information de structure (titre, mots-clés, résumé, bibliographie) dans l'étape d'agrégation des résultats de recherche.

- ***Organisation de mémoire***

Nous structurons ce présent mémoire en quatre chapitres.

-**Chapitre 1 : Recherche d'information agrégée**. Il sera question dans ce premier chapitre de présenter ce paradigme de recherche d'information que nous envisageons d'exploiter dans notre travail. Nous donnerons sa définition, les problématiques qu'il soulève ainsi que les différentes techniques d'agrégation développées dans ce cadre.

-**Chapitre 2 : Documentation scientifique**. Il sera question dans ce chapitre de présenter l'objet de notre recherche qui est la documentation scientifique. Nous donnons sa définition, ses types, ses critères d'évaluation et surtout, on s'intéressera plus à sa structure.

-**Chapitre 3 : Conception du système**, ici nous nous baserons essentiellement sur la technique de Clustering comme technique d'agrégation. Nous allons présenter ce concept, les différents algorithmes utilisées. Nous présenterons ensuite nos choix de conception.

-**Chapitre 4 : Implémentation et évaluation** : Dans ce dernier chapitre, nous présentons nos choix d'implémentation ainsi que l'évaluation de notre solution .

Chapitre 1

Recherche d'information agrégée

1.1 Introduction

Comme mentionné dans l'introduction générale, nous nous baserons sur le paradigme de la Recherche d'information Agrégée pour le développement de notre système. Nous allons donc entamé dans ce premier chapitre la présentation de ce nouveau paradigme de recherche d'information.

L'expansion du Web depuis la fin des années 90 a modifié en profondeur le fonctionnement des Systèmes de Recherche d'Information (SRI. Du côté des documents, des données de différentes formats : sons, images et vidéos, actualités. . . sont venues s'ajouter au format texte traditionnellement traité par les SRI.

Du côté des requêtes, la requête booléenne des débuts de la RI a laissé place dans les années 2000 à des requêtes courtes formées de 2-3 mots-clés dont les moteurs de recherche Web devaient se satisfaire pour répondre à l'utilisateur.

Enfin du côté des modèles et présentation des résultats, les moteurs de recherche Web, après avoir longtemps proposé aux utilisateurs les fameux « 10 liens bleus » (10 bleu links) en réponse à leur requête, incluent maintenant dans leurs pages de résultats des images, des vidéos ou encore des actualités. Lorsque la requête est une entité, les informations liées peuvent aussi être présentées dans un cadre séparé. L'utilisateur est placé au centre de la recherche, son contexte (et son profil lorsque disponible) permettant de fournir des résultats personnalisés L'idée n'est plus de restituer des documents relatifs à une requête,

mais de donner directement à l'utilisateur un aperçu global de l'information liée à son besoin.

C'est autour de ces évolutions, que la recherche d'information agrégée a été définie dès 2008, avec pour but de chercher et d'assembler dans une seule interface de l'information utile provenant d'une ou plusieurs sources. La quantité d'information disponible étant immense, le but est d'en faire le tri et de présenter à l'utilisateur un résultat « résumé » de son besoin. Les limitations de la traditionnelle liste de documents en réponse à une requête s'en trouvent largement atténuées : l'information pertinente n'est plus dispersée dans plusieurs documents, et le résultat présenté se focalise sur le besoin utilisateur (Kopliku et al., 2014)[3].

1.2 La recherche agrégee

L'expression recherche agrégée (Aggregated Search) a été définie pour la première fois dans un atelier à SIGIR , c'est une tâche cherchant à rassembler des informations provenant de sources différentes, et à les présenter dans une seule interface. En d'autres termes, la recherche agrégée tente d'identifier le contenu nécessaire, de l'organiser et de le présenter à l'utilisateur de manière à faciliter sa recherche d'information. Un moteur de recherche agrégé est construit en surcouche d'un ou plusieurs autres moteurs de recherche (sources). Des informations de différents types (image, vidéo, etc.) et de différentes granularités (passage de texte, entités, attributs, etc.) sont reliées et parfois même combinées par une ou plusieurs relations logiques (association, groupe, ordre, etc.) afin de composer un résultat agrégé (Kopliku, 2009).[1]
Contrairement a la recherché d'informations classique qu'est un domaine historiquement lié aux sciences de l'information et à la bibliothéconomie qui ont toujours eu le souci d'établir des représentations des documents dans le but d'en récupérer des informations à travers la construction d'index. L'informatique a permis le développement d'outils pour traiter l'information et établir la représentation des documents au moment de leur indexation, ainsi que pour rechercher l'information.[1] La recherche d'information agrégée peut être décrite selon le schéma de la figure ci-dessous).

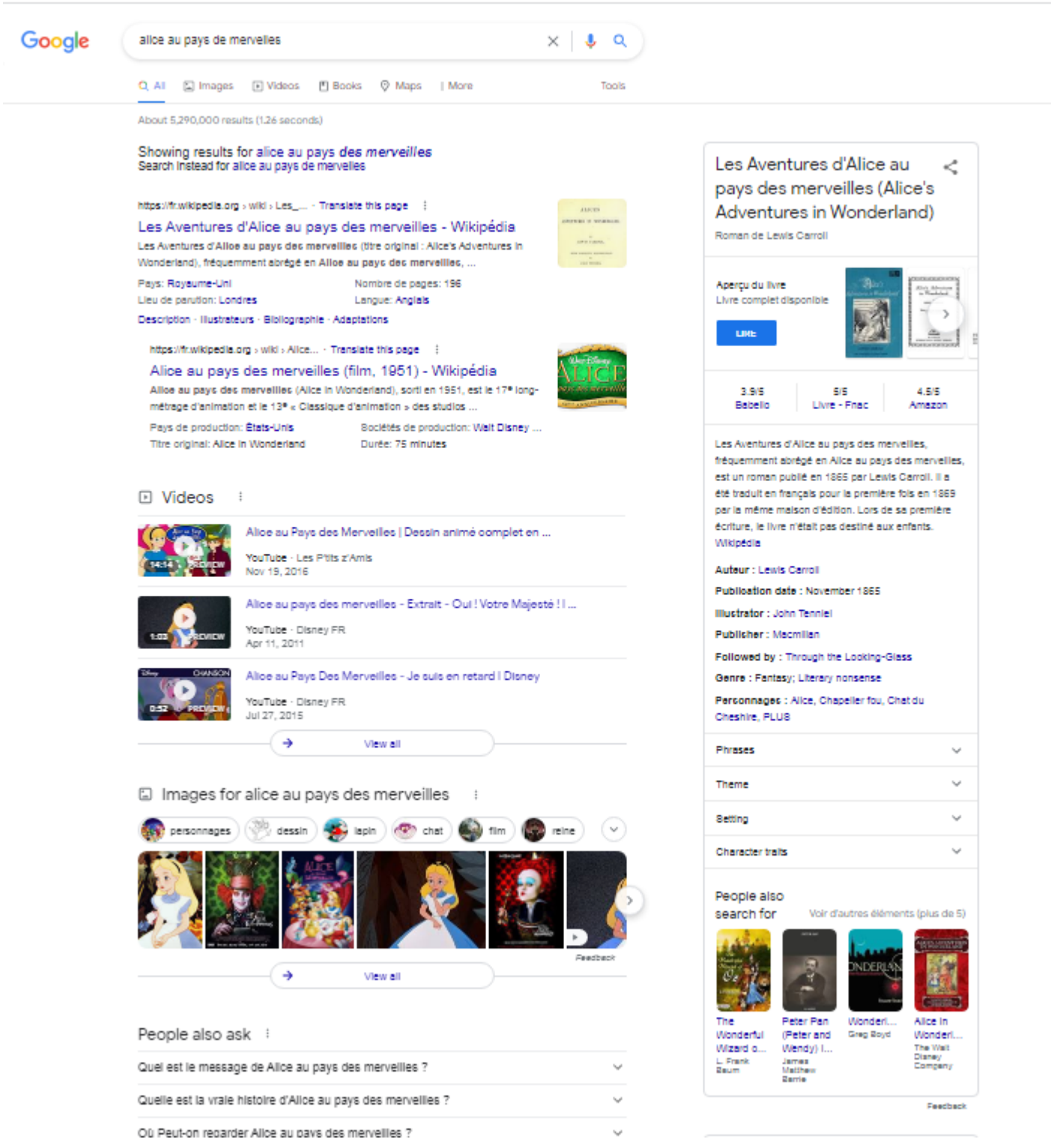

FIGURE 1.1 – Exemple de recherche agrégée-Universel Search[2]

1.3 Les problématiques soulvées par la RI agrégée

Comme on l'a mentionné au niveau de l'introduction, ce nouveau paradigme de la recherche d'information a soulevé un certain nombre de problématiques supplémentaire par rapport aux problématiques classique de la RI agrégée.

1.3.1 Représentation des sources

Les sources ont une certaine représentation interne dans les systèmes de recherche agrégée. Cette représentation peut être aussi simple qu'une description textuelle de la source, mais en général, il contient des termes et des caractéristiques représentatifs de la source extraits de l'échantillonnage ou d'autres techniques d'extraction. [3]

1.3.2 Sélection des sources

La sélection de sources est l'un des problèmes les plus connus de la recherche agrégée avec plusieurs sources. Etant donné un ensemble de sources, son but est de sélectionner les sources susceptibles de répondre à la requête. L'identification des termes clés peut par exemple être utilisée pour sélectionner les sources. La présence de mots-clés spéciaux est souvent utile pour comprendre le type de réponse que l'utilisateur attend. Par exemple, la présence de mots tels que "météo" ou "définition" est un indice utile qui est utilisé par les principaux moteurs de recherche pour déclencher une réponse personnalisée.[3]

1.3.3 Agrégation des résultats

L'agrégation est au cœur de tout système de recherche agrégée. Il n'est plus question de s'arrêter à identifier les éléments d'information pertinents à partir des différentes sources, mais il faut ensuite les agréger et les organiser dans une seule interface de résultats . Nous précisons ici que c'est cette dernière problématique qui nous intéresse le plus dans notre cas.

1.4 Processus génerique de la RI agrégée

On connait le processus classique de la recherché d'informations, dit processus en U, qui relie les requîtes d'une part et les collections de document d'autre part. Ce processus

s'arrête lorsque les résultats pertinents sont sélectionnés. Le processus de la recherché d'informations agrégée ne doit pas s'arrêter là, mais doit également inclure des étapes d'agrégation et de présentation des résultats provenant de différentes sources. Arlind kopliku (kopliku, 2011)[1] propose un processus général d'agrégation des RI Agrégée comprenant les étapes suivantes :

- **Etape 1 :le dispatching de la requête** est l'étape précédant la recherche proprement dite. Il s'agit d'interpréter correctement la requête, de cerner l'intention utilisateur, de reformuler éventuellement le besoin et de décider quelles sources d'information interroger. [4]
- **Etape 2 : la recherche de granules documentaires** se charge d'identifier l'information potentiellement pertinente. Chaque source, grâce à son algorithme de recherche, va renvoyer un ensemble de granules (avec éventuellement un score de pertinence associé). Il est possible d'obtenir en résultat des documents entiers, des parties de documents, ou encore des contenus multimédia issus de moteurs de recherche verticaux. [4]
- **Etape 3 : l'agrégation des résultats** cherche à assembler les différents granules documentaires afin de former le résultat final. Ce dernier, de façon idéale, devra refléter la diversité des résultats, répondre de façon exhaustive à la requête, et ne pas contenir de résultats redondants. Pour ce faire, différentes actions pourront être menées sur les granules : tri, regroupement, découpage en granules plus petits, extraction d'information, etc. [4]

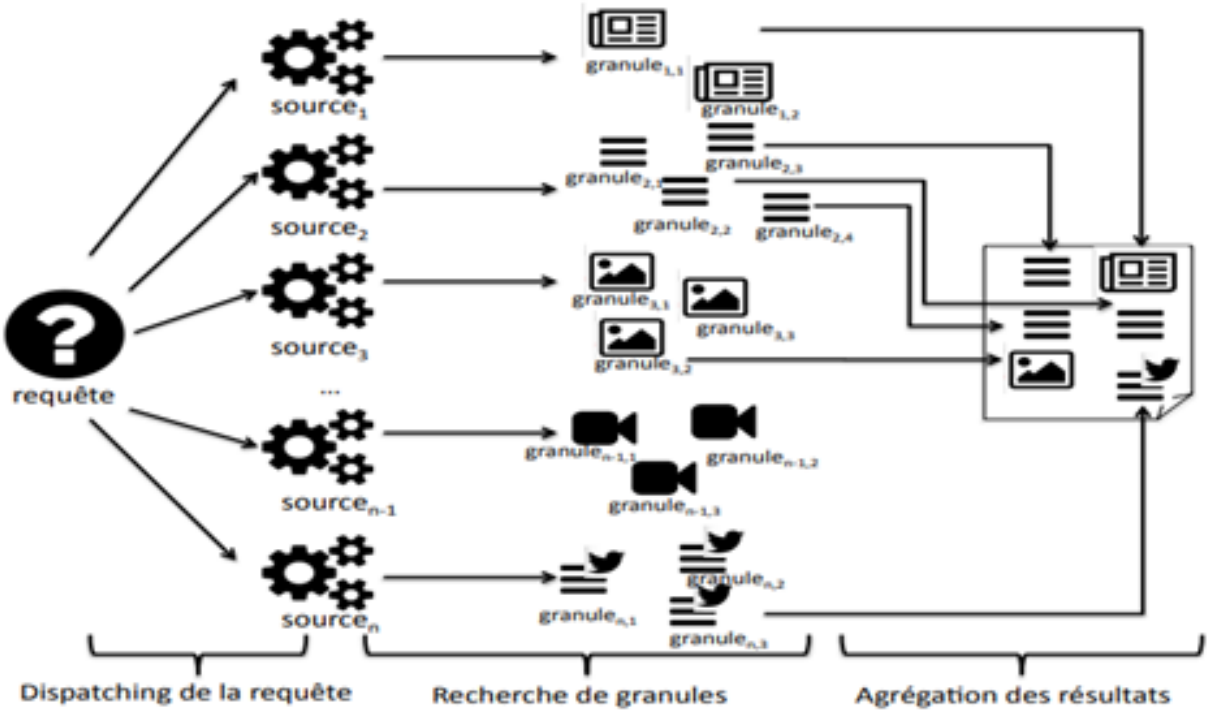

FIGURE 1.2 – Processus de recherche d'information agrégée[4]

1.5 Structure d'agrégat

Dans la RI traditionnelle la réponse à une requête est une liste de documents. Dans la recherche agrégée il devrait être possible d'ajouter l'information de la structure d'agrégat à la requête. La recherche agrégée doit traiter l'organisation du contenu pertinent. Il n'est pas suffisant de collecter (trouver) les informations mais il faut également les organiser pour la visualisation finale. La structure d'agrégat est définie comme toute l'information qui décrit le contenu, l'ordre de visualisation, et les préférences dans la visualisation du document agrégé [5]. Les exemples suivants illustrent des structures partielles de certaines informations agrégées :

- 3 images, 2 vidéos
- Une liste de 4 news, 3 informations supplémentaires.
- Paragraphe A, paragraphe B, paragraphe C, avec l'ordre de visualisation : A, B, C.

1.6 Les approches d'agrégation

Dans cette section, nous passons en revue les travaux directement liés à l'étape d'agrégation des résultats. Plusieurs approches d'agrégation des résultats de recherche ont été développées. Elles sont de deux types : les techniques d'agrégation par génération et celle de fusion :

1.6.1 Approche par géneration

Les approches de génération consistent à carrément générer une réponse à partir des différents éléments pertinents sélectionnés. La réponse agrégée peut être un document généré automatiquement à partir des différents granules d'information pertinents. On trouve dans ce types d'approches les techniques suivantes :

1.6.1.1 Génération du langage naturel

La génération de langage naturel cherche à générer des réponses cohérentes et compréhensibles, formulées en langage naturel, en réponse à un besoin d'information généralement exprimé sous forme de question. Parmi elles nous pouvons citer l'approche de Paris et al. (2010) dans laquelle les auteurs se sont appuyés sur des techniques de traitement de langue pour agréger des résultats de recherche afin de construire un rapport de surveillance, planifier un programme pour des voyages ou produire une brochure pour des organisations. D'autres auteurs comme (Sauper, Barzilay, 2009) ont cherché à construire automatiquement des articles médicaux pour Wikipedia à partir de modèles qu'ils ont eux-mêmes générés. Ils ont montré que l'intégration d'informations structurées provenant d'articles déjà existants dans Wikipedia dans le processus d'apprentissage d'extracteurs de contenus améliore la qualité des articles construits. Les résultats de l'évaluation confirment les avantages d'intégrer des informations structurées dans le processus de sélection des contenus par rapport aux approches qui n'ont pas un modèle de structure pour chaque thème vu.[1]

1.6.1.2 Systèmes de Question-Réponse

La réponse aux questions (QA) diffère de la recherche classée traditionnelle car elle ne vise pas une liste de documents, mais une ou plusieurs réponses. Il représente un cas d'étude intéressant pour la recherche agrégée car ces réponses peuvent ne pas exister, il

faut les produire, les extraire et les assembler . Dans QA, on retrouve également les trois composants principaux du processus de la RI agrégée. En ce qui concerne la répartition des requêtes, les requêtes dans QA ne sont pas du texte libre plutôt que des requêtes de type question. Pour les questions, il existe différentes taxonomies et il est dans l'intérêt du système d'AQ de comprendre le type de requête. Nous pouvons énumérer ici quelques types de questions bien connues telles que les questions « Qui », « Quoi », « Où », « Quand » ou « Oui/Non ». De plus, les approches existantes recherchent des entités nommées et d'autres faits utiles dans la requête, par ex. "Où en Afrique les scientifiques envoyés par Napoléon ont-ils trouvé Rosetta Stone ?" Cette question contient différentes entités nommées (Napoléon, Afrique...) et faits (en Afrique, scientifiques). Des études de cas et des observations intéressantes sur la relation entre la recherche d'agrégation.[3]

1.6.1.3 Agrégation relationnelle

La recherche agrégée relationnelle est une généralisation de la recherche relationnelle et de la recherche d'entités (Balog et al. 2009). Elle peut être vue comme un paradigme de recherche basé sur les relations entre les différents nuggets. En plus de la recherche de nuggets, la recherche agrégée relationnelle doit trouver des relations, nécessaires ensuite pour la phase d'agrégation des résultats.[3]

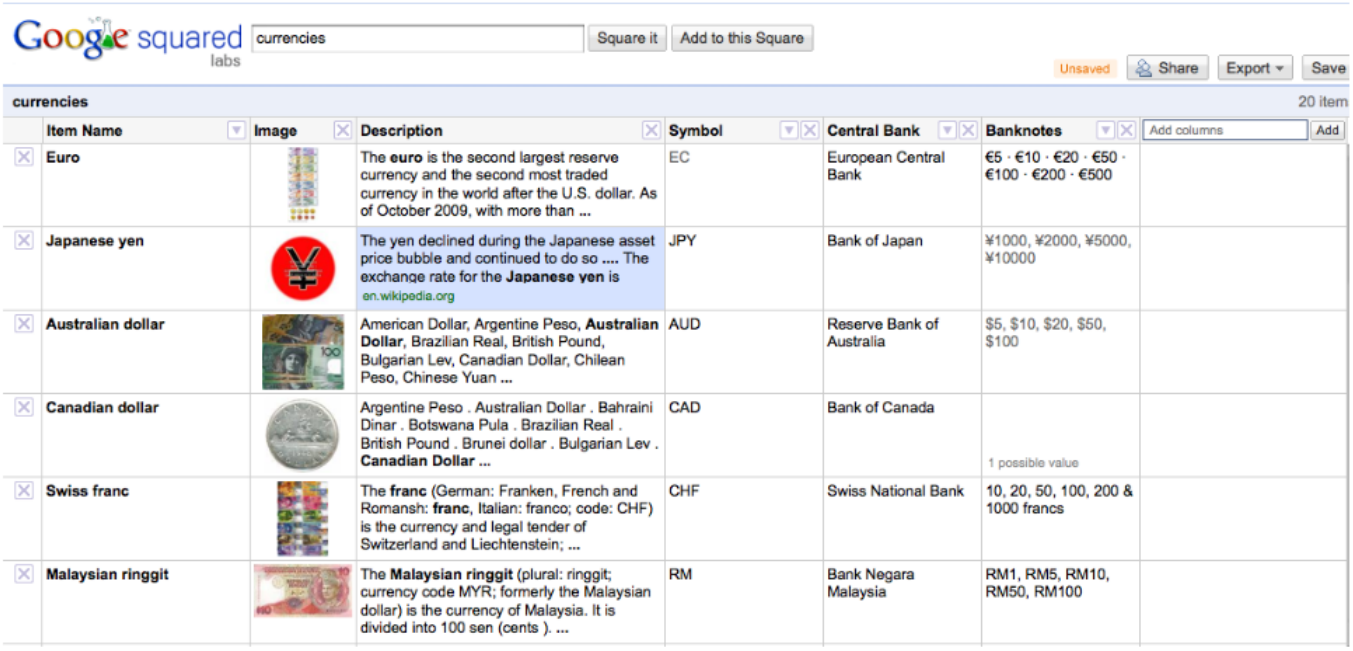

FIGURE 1.3 – Exemple de résultats de Google Squared

1.6.2 Approches de fusion

Les approches de fusion, contrairement à celle de génération ne consistent pas a générer (créer) des réponses, mais à trouver un moyen de fusionner les différents éléments pertinents sélectionnés à partir de sources différentes et des les organiser dan la même page de résultats. Parmi les techniques issue de cette approche, on peut citer les suivantes :

1.6.2.1 La recherche agrégée INTER-VERTICALE(CROSS VERTICAL SEARCH)

Le principe de la recherche agrégée inter-verticale est d'agréger des résultats de recherche provenant de différents moteurs de recherche verticaux, la plupart du temps dans un contexte de recherche web. Les différentes recherches verticales renvoient des informations d'un seul type de support (image, vidéo, etc.) ou de contenu (news, livre, etc.). Ces informations sont ensuite triées et agrégées pour être présentées à l'utilisateur.
L'agrégation de différentes recherches verticales est l'interprétation la plus commune et la plus évidente de la recherche agrégée. En raison de l'augmentation des différents types de support et de contenu d'information dans la recherche d'information, il est devenu inévitable de les intégrer dans une même interface pour pouvoir les exploiter d'avantage. La majorité des moteurs de recherche adoptent déjà ce type d'agrégation. Parmi eux, nous pouvons citer Google Universel 1 (voir exemple sur la figure 3), ou encore Bing .[3]

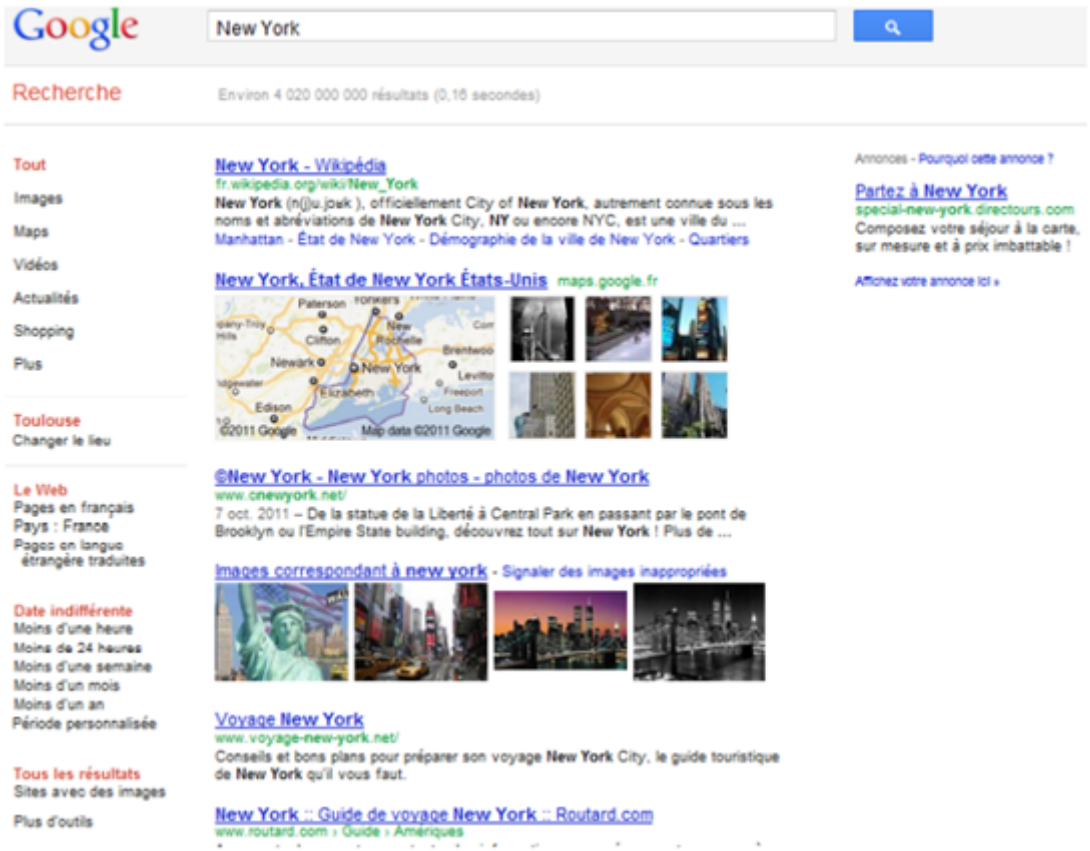

FIGURE 1.4 – Résultats de la requête New York pour le moteur de recherche Google Universal.[1]

1.6.2.2 Approche d'agrégation par clustering

Cette approche consiste à regrouper les documents après récupération sous forme de clusters, et présenter un résumé de chacun de sorte que l'utilisateur peut choisir son groupe d'intérêt. Cette approche été proposée par Zeng et al qui considèrent que le regroupement des résultats de recherche dans des clusters permet d'avoir des documents qui se concentrent sur certains aspects de la requête. Parmi les systèmes basés sur cette technique, nous trouvons Clusty, QCS (Query, Cluster, Summarize) qui effectue les tâches suivantes en réponse à une requête :

- Récupère les documents pertinents.
- Sépare les documents récupérés en groupes par sujet.
- Crée un résumé pour chaque cluster.

D'autres exemples de systèmes de recherche d'information employant des algorithmes de Clustering pour organiser les ensembles des documents récupérés comprennent : Velocity/Clusty (Vivisimo, 2006), Infonetware / RealTerm (Infogistics, 2001), WiseNut (LookSmart,2006), Accumo (Accumo, 2006), iBoogie (CyberTavern, 2006), et le KartOO et systèmes UJIKO (KartOO, 2006). Ces systèmes organisent les documents en clusters et génèrent une liste de mots clés associés à chaque cluster. Les deux derniers systèmes présentent également des représentations graphiques des clusters résultants. Comme avec le système de recherche ci-dessus, ces systèmes présentent aussi des extractions de document contenant un ou plusieurs termes de la requête, mais le seul résumé présenté est la liste de mots-clés.[6]

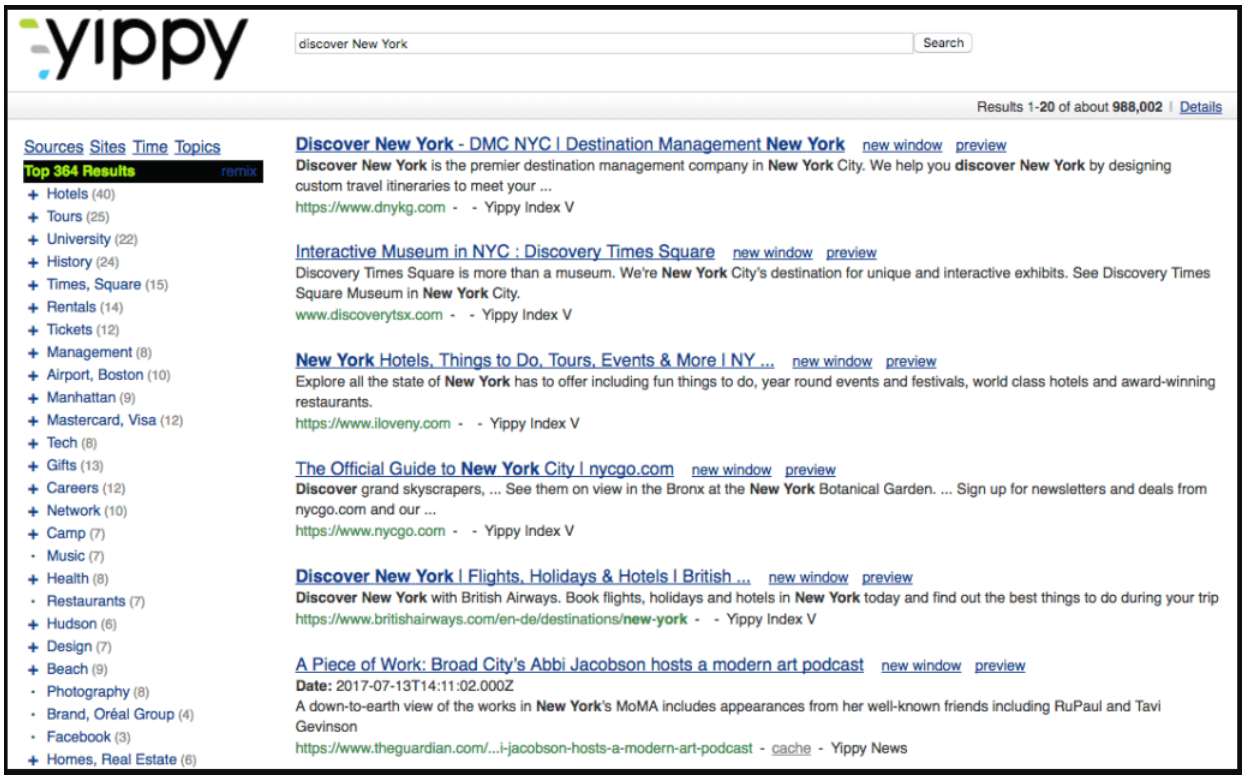

FIGURE 1.5 – Exemple de résultats de recherche du moteur Clustry/yippy

1.7 Conclusion

On a résumé dans ce chapitre brièvement la recherche d'information agrégée et son processus ou on a motionné la définition de la recherche agrégée ainsi que ses problématiques et ses approches.

Chapitre 2

Documentation scientifique

2.1 Introduction

Au sein de la communauté scientifique, l'information passe essentiellement par le biais des publications scientifiques. Ces publications occupent aujourd'hui une place primordiale dans la recherche. Elles constituent l'objectif même de la recherche scientifique étant donné qu'un chercheur est généralement évalué par ses publications. De ce fait et vu l'importance de ces publications, la communauté scientifique doit essayer d'uniformiser ses publications pour qu'elles soient facilement exploitables par tous ses membres n'importe où dans le monde sans obstacles linguistiques, conceptuels et de normalisation.

2.2 C'est quoi un article scientifique

L'article scientifique est un texte académique très utile dans les études collégiales et universitaires. Il sert à informer le lecteur qui s'intéresse activement à un domaine en particulier. Les auteurs (chercheurs) sont des spécialistes et des professionnels du domaine. L'article scientifique est publié dans un périodique (revue) spécialisé du domaine et il est arbitré, évalué et révisé par un comité de lecture (pairs) formé d'experts et de spécialistes du domaine.[7]

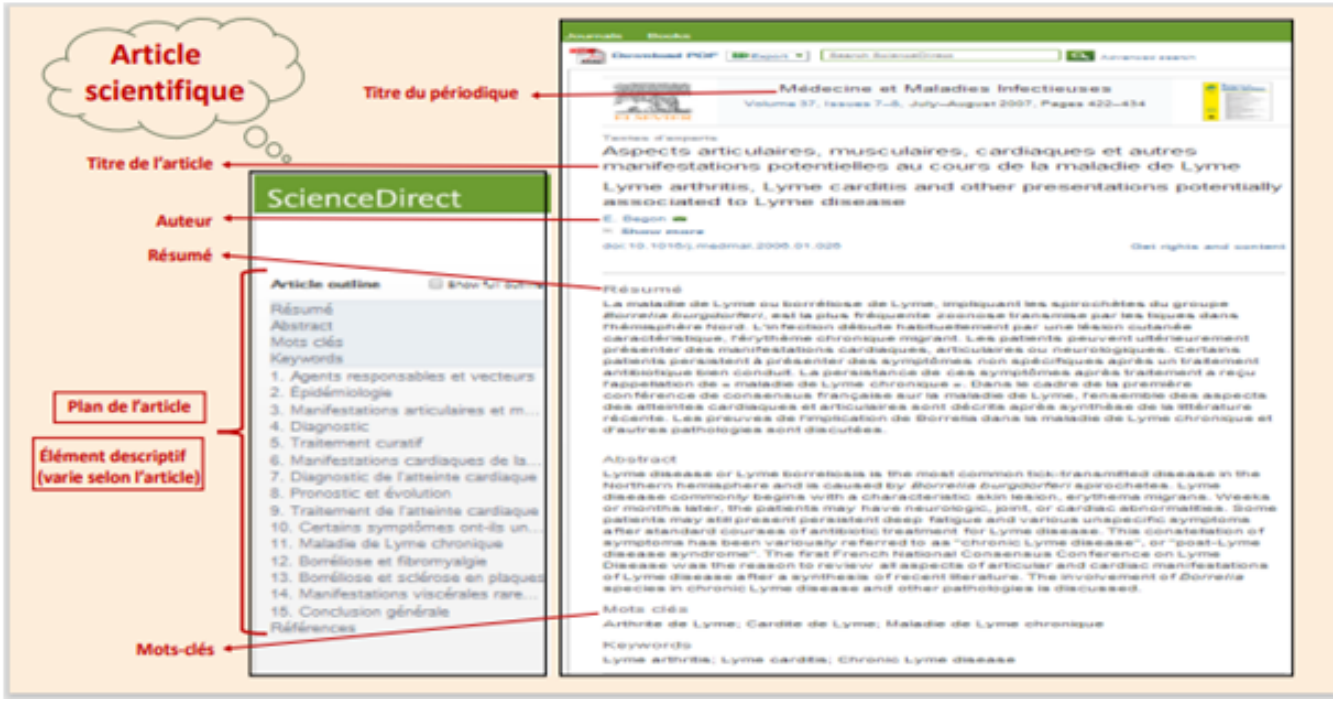

FIGURE 2.1 – Exemple d'un article scientifique[7]

2.3 Comment identifier les articles scientifiques

Critères	**Description**	**Exemples**
Auteur(s)	- Chercheur(s) universitaires(ph.D). -Scientifique(s) d'instituts, de chaires ou de centres de recherche.	-Chaire en entrepreneuriat et innovation. -Faculté des sciences de l'administration. -Département des relations industrielle.
Editeur	- Éditeurs scientifiques. -Département et faculté universitaire.	-Elsevier. -Département de relation industrielles de l'U.
Revue	-Révisé par les paires /Comité de lecteur.	-Mentionné au début de la revue.
Forme	-Généralement plus de 10 pages. -Peut inclure des tableaux(rarement des photos.	
Contenu	1.Résumé(abstract). 2.Mots-clés. 3.Introduction. 4.Cadre théorique(et/ou hypothèses). 5.Méthodologie de recherche. 6.Résultats de recherche. 7.Discusion des résultats. 8.Conclusion. 9.Bibliographie et offée(références).	

TABLE 2.1 – Tableau décrivant critères article scientifique[8]

2.4 Les types des écrits scientifiques

La littérature scientifique constitue un ensemble très varié de documents :

2.4.1 Les journaux scientifiques

Appelés encore les revues scientifiques, elles sont définies par DEVILLARD MARCO (1993) comme suit : "une publication en série, à parution régulière, dotée d'un titre déposé et composée d'une suite d'articles évalués par un comité de lecture en fonction de critères scientifiques ".[9]

2.4.2 L'article scientifique proprement dit ou encore le "document scientifique"

C'est celui qui publie les résultats originaux d'une recherche. Dans sa thèse DEVILLARD (1991) le représente : "outre le fait qu'il représente pour les chercheurs le principal moyen d'expression, il est aussi le moyen de communication le plus commode et le plus utilisé entre les différent membres d'une même communauté scientifique ". Quand à CROOKES (1986), il donne la définition suivante : " Un document scientifique peut être défini comme un type d'écrit scientifique, basé sur la simple investigation dont le but est de contribuer au progrès de la science ou de la technologie ".[9]

2.4.3 La revue génerale ou encore " REVIEW PAPER "

Ce type d'article ne contient pas les résultats originaux d'une recherche et donc n'est pas considérée comme publication primaire. Ces types d'articles peuvent contenir des nouvelles informations qui n'apparaissent pas dans le document original de la recherche. Cependant le but de ce type d'article est de réviser et critiquer la littérature précédemment publiée et la mettre dans une certaine perspective (DAY, 1989).[9]

2.4.4 Le rapport de conférence

Il présente une ou plusieurs interventions ainsi que les discussions dans une conférence entre scientifiques que ce soit un congrès, séminaire ou autres.[9]

2.5 La structure des écrits scientifiques

2.5.1 La structure physique

L'écriture scientifique répond à certaines exigences physiques, qui varient d'un article à l'autre Soutien (revues, livres, articles, etc.) et discipline. Quelques éléments de structure physique à connaître :[10]

- Mise en page : Pleine page, Colonnes, Marges,...
- Caractères : polices, typographie,...
- Taille du fichier : format de page (A4 ou autre), taille...
- Le volume du document : le nombre de pages, le nombre de mots...
- Sur un côté ou les deux côtés du papier, espacés (simple, double...)

2.5.2 La structure logique

L'écriture scientifique se conforme au fait que la documentation scientifique doit avoir une structure logique bien définie et claire. En passant en revue certaines littératures scientifiques, nous avons constaté que la structure logique de ces littératures est généralement résumée dans un plan ou un catalogue. En effet, pour permettre aux chercheurs d'accéder et de comprendre l'écriture scientifique Les scientifiques doivent organiser leur travail de manière assez claire. BENICHOUX déclare : « La communauté scientifique doit œuvrer pour lever les barrières au transfert de la science et la rendre accessible à l'international, Les éléments les plus importantes concernent l'ordre et la séquence des présentations Scientifique. La plupart des articles scientifiques sont structurés selon le plan **IMRaD** : **I**ntroduction, **M**éthodes ,**R**ésultats et (and) **D**iscussion. Le lecteur, habitué à cette structure, y retrouve facilement ses marques et sait rapidement trouver ce qu'il cherche. Sa grande lisibilité et son caractère quasi universel en font une excellente base pour la structure d'une publication scientifique. En français on dit IMRED pour : Introduction Méthode Resultats Et Discussion. Il est à noter que cette structure varie selon le type d'écriture (article, mémoire, ...) et la discipline. En fait, il est loin de standardiser toute la littérature scientifique.[11]

2.5.2.1 L'article de recherche scientifique

Le plan IMRED : Les principaux articles, notamment ceux en sciences expérimentales (biologie, médecine, agronomie, etc.), sont généralement structurés selon le programme IMRED : les différentes parties du programme sont :

Introduction : Dans cette partie, l'auteur de l'article doit dire les points principaux et il doit citer les travaux de certains des auteurs répertoriés dans la liste de référence pour se positionner.[11]

Matériels et méthodes : Le but de cette section est de permettre aux auteurs de l'étude (lecteurs et évaluateurs de l'article) de reproduire, si nécessaire, tous les détails possibles du travail qu'ils ont effectué pour leur validation.[11]

Résultats : Dans cette section, les résultats obtenus à partir des expériences sont présentés en détail. Habituellement, cette section contient des tableaux, des graphiques... pour rendre la lecture plus claire et plus facile.[11]

Discussion : Ce chapitre retient des commentaires sur les résultats, qui sont présentés sous forme d'unité unique ou de sous-unités multiples, en comparant les résultats entre eux, en les comparant avec des résultats publiés, et enfin en répondant à l'hypothèse de travail présentée dans l'introduction et les détails. sont décrites dans la section Matériels et méthodes.[11]

Les clés du texte : En plus des parties essentielles de (décrites dans le plan IMRED ou autrement), en ce qui concerne la structure et le contenu textuel des articles scientifiques, il existe d'autres éléments de référence qui jouent un rôle important selon le type d'article. [11]Ces éléments sont appelés clés textuelles et comprennent :

-Le titre : Benissieux a dit : « le titre d'un article scientifique sert d'enseigne , et le résumé est une vitrine, vous devez donc choisir avec soin.[11]

- **L'auteur** : Trouvez généralement le nom de l'auteur et l'affiliation de l'institution où le sujet de recherche de l'article est mené.[11]

- **Le résumé** Il est généralement placé au début d'un article et titue la partie la plus lue d'un article scientifique et doit être écrit avec soin.[11]

Les mots clés : Spécifique aux articles scientifiques. Ces mots sont généralement choisis par l'auteur de l'article.[11]

la bibliographie : Les articles scientifiques se caractérisent par de solides bibliographies.[11]

Quant aux autres travaux scientifiques, ils ont généralement une structure logique peu claire.[11].Cette structure est généralement définie dans un répertoire ou répertoire de documents,image ci-dessous défini cette structure

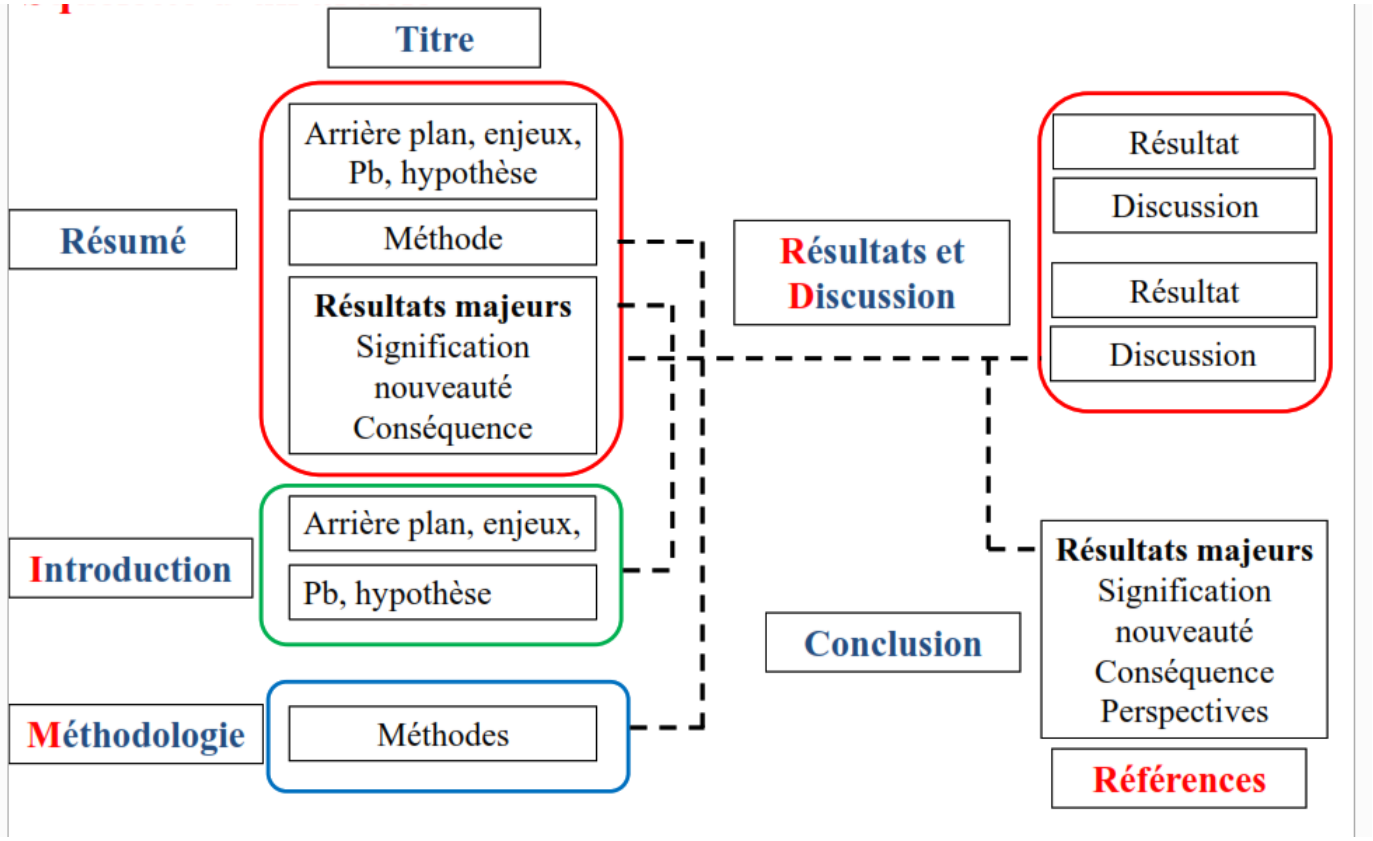

FIGURE 2.2 – squelette d'article scientifique[12]

2.5.2.2 Mémoires et thèses

Le travail académique (thèse et mémoire) n'a pas de structure logique standardisée Surtout dans le texte principal, ces structures varient d'une discipline à l'autre. nous présentons ici la structure issue des travaux universitaires proposés par ROOVEYRAN :

- Préliminaire (couverture, page de titre, résumé, etc.) ;
- Texte (introduction, corps, conclusion...) ;
- Références (bibliographie, annexes, index) ;
- Tableaux (tableaux matériaux, illustrations...) ;
- enfin le résumé et les mots-clés..[11]

2.5.2.3 Les ouvrages scientifiques

En général, la structure logique d'un livre varie d'une discipline à l'autre et même d'un livre à l'autre. Selon FEBVRE GIOROAN : « Le travail scientifique est structuré. Les subdivisions qui lui sont propres sont :

• Avant-propos et remerciements,

• table des matières ou résumé,

• introduction,

• bibliographie,

• glossaire... I

Ils ont un agencement et un contenu, qui aide à la recherche d'informations dans des livres parfois volumineux et complexes".[11]

2.6 Médiums de publication d'articles scientifiques

Nom	**Description**
-Atelier(workshop) dans conférence scientifique.	-Présentation d'une idée originale,mais incomplète.
-Conférences scientifiques.	-Présentation d'une idée nouvelle et originale.
-Journaux scientifiques.	-Approfondir une idée originale ou récente(pouvant etre avoir déjà été dans une conférence). -Processus itératif.
-Recueil d'articles(chapitres de livre).	-Contient des articles (chapitres) sur un thème ciblé.
-Livre de référence(#article).	-Présentation uniforme.

TABLE 2.2 – Tableau décrivant médiums de publication articles scientifiques[13]

2.7 Exemples de moteurs de recherche scientifique

2.7.1 Google scholar

FIGURE 2.3 – Google scholar[14]

Google Scholar est, comme vous pouvez déjà l'imaginer, un produit de la firme Google. Il a été lancé en 2004 et permet aux universitaires de faire des recherches sur des articles scientifiques qui sont approuvés ou non pas les comités de relecture. Contrairement à son aîné qui interroge tout le web pour proposer des réponses à des recherches, Google Scholar quant à lui se concentre strictement sur des thèses, des références bibliographiques à caractère académique, des livres scientifiques ou des citations.[14]

2.7.2 Dimensions

FIGURE 2.4 – Dimensions[14]

Dimensions est un moteur de recherche lancé en 2018 par la société commerciale Digitale Science. Il est comme une base de données bibliographique multidisciplinaire. Ce moteur de recherche est disponible en deux versions : une version allégée qui est accessible gratuitement et une version payante. Malgré son jeune âge, il ambitionne véritablement de concurrencer certains mastodontes en matière de bases de données multidisciplinaires tels que Scopus d'Elsevier et ses 70 millions de références et Web of Science Core Collection de Clarivate Analytics et ses plus de 77 millions de références.[14]

2.7.3 Scinapse

FIGURE 2.5 – Scinapse[14]

Scinapse est l'un des derniers moteurs de recherche académique mis sur le marché. Il a été lancé en 2019, par un réseau de chercheurs et développeurs sud-coréens. Ce moteur de recherche se veut être le concurrent numéro 1 de Google, ce qu'il annonce d'ailleurs depuis sa page d'accueil : «Nous sommes meilleurs que Google».
S'il est lancé seulement en 2019, ce moteur de recherche compte déjà plus de 48000 revues, 200 millions d'articles avec plus de 50000 chercheurs inscrits dans plus de 196 pays. Selon ses données, les chercheurs inscrits ont déjà accédé à plus de 50 millions de documents depuis son moteur de recherche. Pour fonctionner, il puise ses données à partir des projets, financements, publications et citations qu'il collecte à partir de ses sources telles que PubMed, Microsoft Acadmic Graph, Springer Nature et Open Research Corpus.[14]

2.7.4 Semantic Scholar

FIGURE 2.6 – Semantic scholar[14]

Semantic Scholar est un moteur de recherche gratuit et à but non lucratif mis sur le marché pour le monde universitaire. Il a été lancé en 2015 par Allen Institute for Art Intellignece de Paul Allen, cofondateur de Microsoft. Ce moteur de recherche académique indexe les documents scientifiques libres accès sur internet avec des liens vers des articles de blog et reportages.[14]

En janvier 2020, ce moteur de recherche a enregistré plus de 180 millions articles qu'il puise dans des dizaines de sources telles que Springer Nature, ArXiv, Wolters Kluver, De Gruyter pour ne citer que celles-là. Comme Google Scholar, Semantic Scholar permet également de filtrer les résultats de recherche par période de publication, par un ou plusieurs mots ou expression.[14]

2.7.5 Connected papers

on peut utiliser ce moteur pour :

— Entrez un papier typique et nous vous construirons un graphique des papiers similaires dans le domaine

— Explorez et créez plus de graphiques pour les articles intéressants que vous trouvez - vous aurez bientôt une véritable compréhension visuelle des tendances, des travaux populaires et de la dynamique du domaine qui vous intéresse.Créer la bibliographie de votre thèse Commencez par les références que vous voudrez certainement dans votre bibliographie et utilisez Connected Papers pour combler les lacunes et trouver le reste !

2.8 Conclusion

Nous avons donné dans ce chapitre un bref aperçu sur les publications scientifiques. Nous avons défini quelques éléments caractéristiques d'un article scientifique ainsi que sa structure physique et logique.

Nous avons ensuite présenté quelques exemples des meilleurs moteurs de recherche scientifiques qui existent actuellement. Il en ressort que malgré la puissance etla grande pertinence de ces moteurs de recherche, l'utilisateur qui les utilise éprouve encore beaucoup de difficultés à bien explorer l'espace de résultats constitués d'un long listing de liens vers des articles potentiellement pertinents. Ces moteurs ne lui fournissent pas assez d'outils pour le guider vers l'information qu'il cherche. Il est souvent obligé de télécharger et les plusieurs articles proposés pour enfin tomber sur l'info recherchée. Cette difficulté est due au fait que ces moteurs ne fournissent pas d'outils d'organisation et d'agrégation des différents résultats de recherche. Ils s'arrêtent à la récupération des articles pertinents des différents sources et de les fusionner dans une seule liste de résultats.

Dans le prochain chapitre, nous présenterons la conception de système nous présenterons essentiellement la technique de clustering comme technique d'agrégation.les différents algorithmes utilisées. Enfin nous présenterons nos choix de conception.

Chapitre 3

Conception du système

3.1 Introduction

Après avoir présenté l'état de l'art des concepts et techniques liés à notre étude de cas dans la partie précédente de ce mémoire, nous allons commencer la phase de conception de notre système dans ce chapitre. Nous passerons en revue les objectifs de notre travail, présenterons les défis que nous avons rencontrés et nos contributions.

Nous rappelons ici que l'objectif principal de notre travail est de développer un système de recherche d'information agrégée pour la recherche d'articles scientifiques, ce qui permettra de :

- Aider le chercheur dans son activité de recherche.
- Guider l'utilisateur vers l'information pertinente.
- Exploiter la spécificité les articles scientifique pour mieux organiser l'espace des résultats.
- Exploiter l'information de structure (titre, mots-clés, résumé, bibliographie) dans l'étape d'agrégation des résultats de recherche.

Comme cela a été mentionné dans le chapitre précédent, malgré la puissance et la grande pertinence des moteurs de recherche scientifique, l'utilisateur qui les utilise éprouve encore beaucoup de difficultés à bien explorer l'espace de résultats constitués d'un long listing de liens vers des articles potentiellement pertinents.

Nous pensons que compléter ces moteurs de recherche par une étape d'agrégation des résultats en exploitant le paradigme de la RI agrégée donnera plus de solution

à l'utilisateur et lui permettra de gagner en temps et en effort nécessaire dans son activité de recherche.
Au lieu de présenter à l'utilisateur un listing d'articles scientifiques susceptibles de satisfaire son besoin, nous envisageons de lui présenter un espace de résultats bien organisé et facile à explorer. Cela passera nécessairement par une étape d'agrégation des différents résultats de recherche via l'une des techniques d'agrégation cités précédemment dans le chapitre 1.

3.2 Quelle technique d'agregation pour notre cas

De toutes les techniques d'agrégation présentées au niveau du chapitre 1, nous estimons que l'agrégation par Clustering est la plus indiquée pour notre cas. En effet, cela est justifié par les constats suivants :
- L'approche par Clustering est la plus indiquée pour mieux organiser un espace de résultats contenant une grande quantité d'information. Ce qui est généralement le cas dans les systèmes de recherche d'information scientifique. Le nombre de résultats est souvent très grand.
- L'approche par Clustering permet de guider l'utilisateur sur les articles en liens avec le concept qui l'intéresse et lui fais gagner du temps et de l'effort à fournir pour arriver à son besoin.
- La structure des articles scientifiques (titre, mot-clés, résumé, nom de revue, références, nom auteur...) nous donne aisément les éléments caractéristiques sur lesquels on peut se baser pour la tâche de regroupement des articles.
- Les autres approches par génération ne sont pas très indiquées étant donné que le chercheur ne cherchent pas une réponse à une question précise.

3.3 Classification automatique de documents

Avant d'aller dans le détail des étapes de notre solution, nous allons rappeler ici brièvement le principe de classification (Clustering) automatique de documents ainsi que les différents concepts y afférant.

3.3.1 Objectifs de la classification automatique de documents

Comme le nombre et le volume des documents numériques s'accroissant de façon exponentielle, on a besoin de les catégoriser afin de facilité leur manipulation. La classification de textes a pour objectif de regrouper les textes similaires, c'est à dire thématiquement proches, au sein d'un meme ensemble. En d'autres termes, trouver un algorithme permettant d'assigner un texte à une classe avec le plus grand taux de réussite possible.

3.3.2 Classification supervisée vs non-supervisée

La classification automatique cherche à répartir un corpus en groupes de documents (catégories, classes, clusters) de façon à mettre ensemble les documents qui se ressemblent et de séparer celles qui différent. Deux méthodes de classification sont utilisées, la méthode supervisée et non supervisée.

• **Classification supervisée :**

C'est une méthode consiste à attribuer une ou plusieurs classes à un document dont les classes sont connues à priori, c'est-à-dire basé sur des documents d'entrée et de sortie étiquetées. C'est une méthode qui permettre à un système d'etre capable de prédire la classe d'un nouveau document non classé (non étiqueté) à base d'un ensemble de descripteurs. La performance de la classification dépend toujours de l'efficacité de la description.

• **Classification non-supervisée (Clustering)** :

Le Clustering est une méthode d'apprentissage non-supervisé permettant de trouver des patterns dans les données. Par contre à la première approche c'est que dans cette approche les classes ne sont pas connues à priori, les documents ne sont pas étiquetés au préalable. Cela signifie un regroupement ou partitionnement des documents en fonction de leurs similarités (regroupement des documents qui se ressemblent). On dispose des documents non classés dans le but de trouver de modèles communs.[15]

3.3.3 Processus de classification automatique de documents

La classification automatique de documents passe par plusieurs étapes :

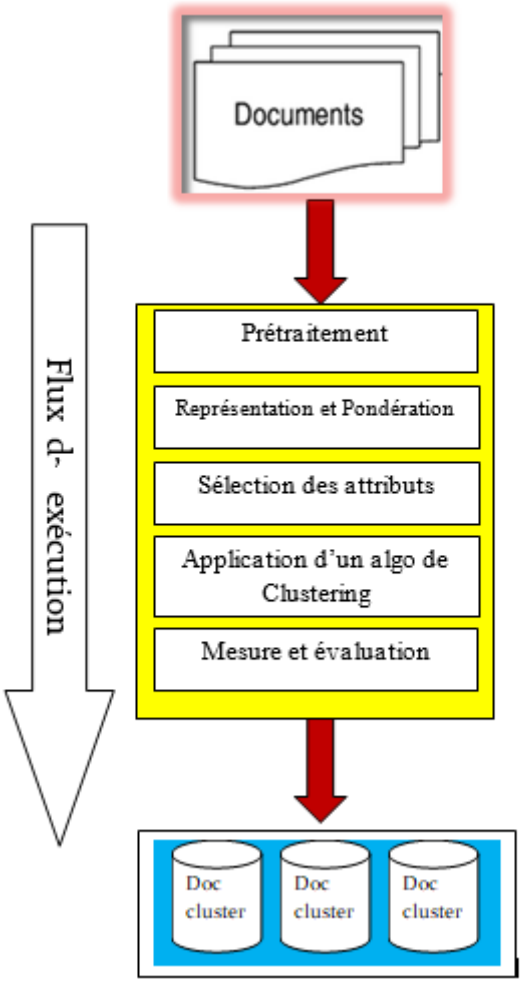

FIGURE 3.1 – Processus de classification automatique de documents[16]

3.4 Architecture générale de notre solution

L'architecture générale de notre système est schématisée dans la figure suivante.

- Tout d'abord la requête utilisateur est envoyée vers un ou plusieurs moteurs de recherche scientifiques.

- Chaque moteur va renvoyer un certain nombre de résultats de recherche.

- Les différents résultats de recherche sont ensuite agrégés via un Clustering pour regrouper ensembles les articles qui concernent les mêmes sujets.

- Un espace de résultats final est ensuite présenté à l'utilisateur sous forme de

résultats regroupé sous formes de cluster. Chaque Cluster renferme des articles qui sont les plus proches entre eux en termes de sémantique.

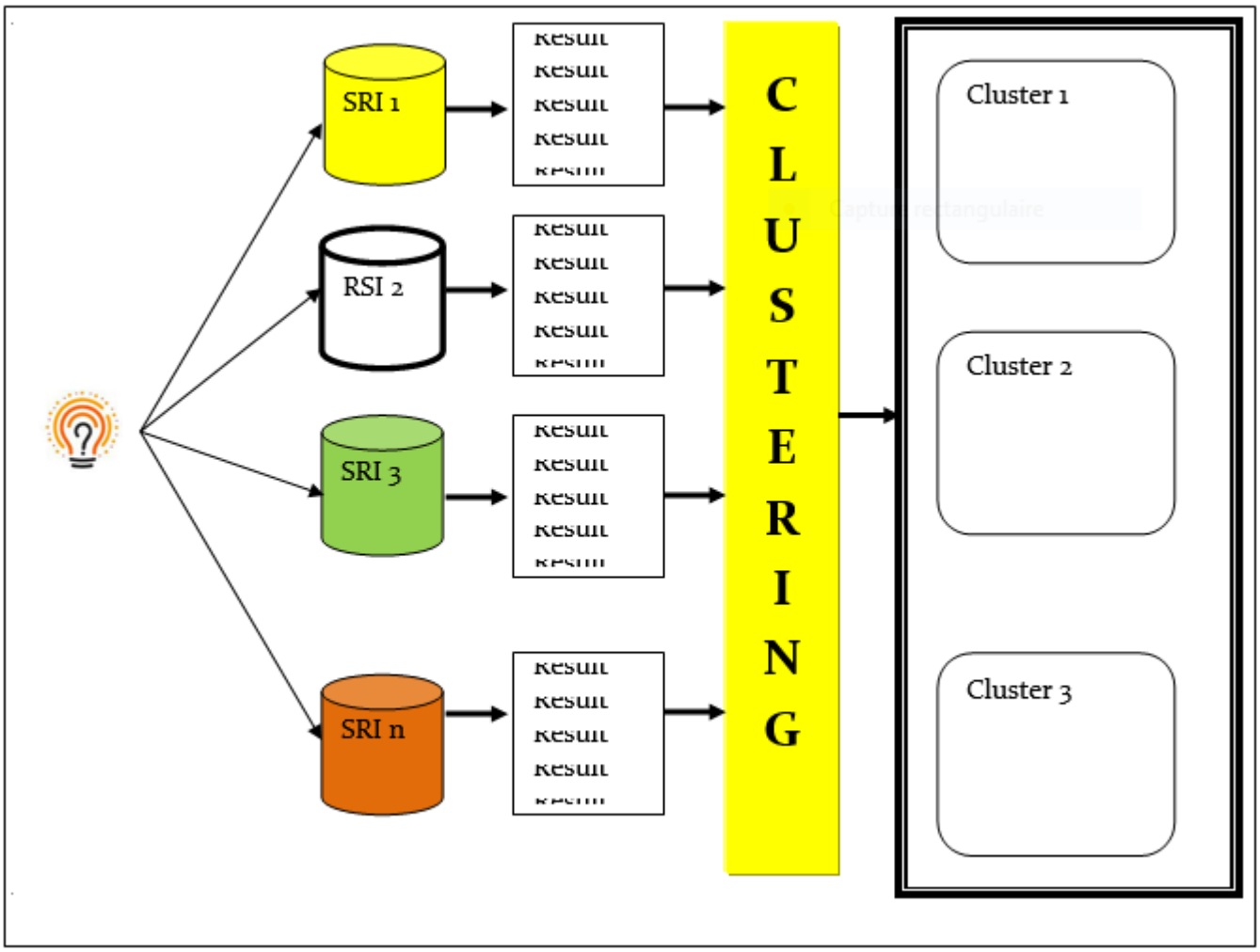

FIGURE 3.2 – Architecture générale du système

Dans ce qui suit nous allons détailler chaque étape de notre solution, essentiellement les étapes liées aux Clustering des articles scientifiques avant leur présentation à l'utilisateur.

3.4.1 Présentation : collection et nettoyage des articles

Les documents en entrée dans notre processus de Clustering sont issus des différents moteurs de recherche scientifiques vers lesquels la requête utilisateur est envoyée. L'ensemble de ces résultats constitue notre corpus de documents à classifier. Les algorithmes de Clustering ne travaillent pas directement sur les documents mais

sur des représentations de ces documents.

Les moteurs de recherche renvoient des fichiers sous format JSON. Pour chaque résultat (article scientifique) nous nous intéressons aux rubriques : Titres, mots-clés, résumé. Ces rubriques contiennent des informations essentielles qui nous permettent de tirer les descripteurs pour chaque article scientifique.

Le résultat de cette étape est d'associer un contenu composé des mots existant dans les rubriques titre, résumé et mots-clés à chaque article scientifique .

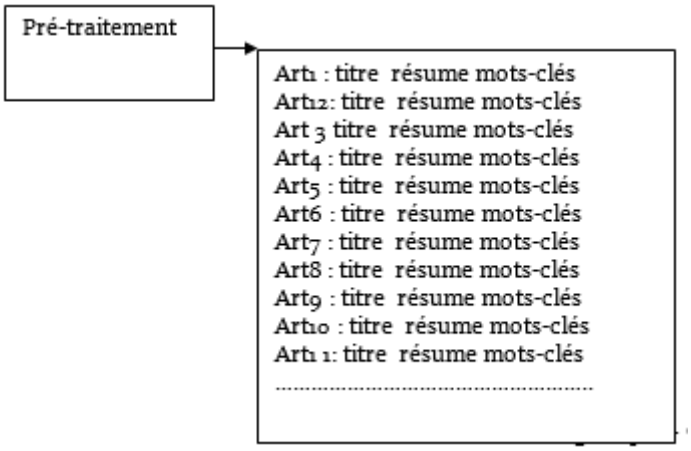

FIGURE 3.3 – Collection et nettoyage articles

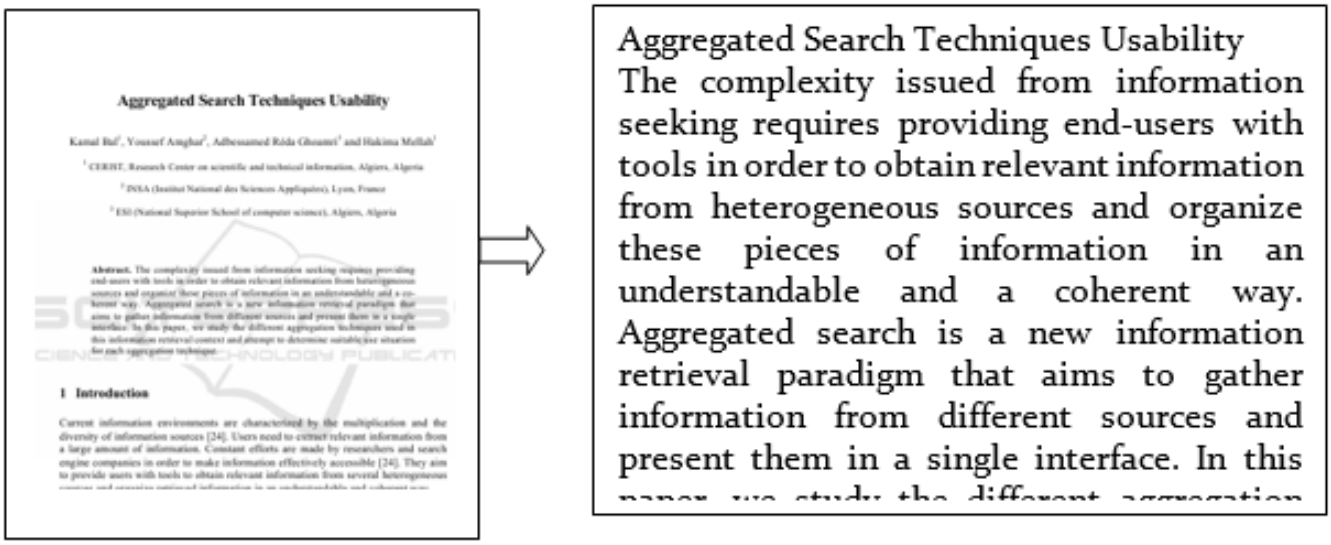

FIGURE 3.4 – Exemple d'un article scientifique

3.4.2 Représentation et pondération des articles

Le but de cette étape est de générer une représentation terminologique sous la forme d'une matrice, où les lignes correspondent aux articles et les colonnes correspondent aux termes. Chaque article (ligne) sera représenté par un vecteur pondéré, et chaque coordonnée correspondra au poids de chaque entrée de l'article. Comme le montre la figure ci-dessous, cela passera par les étapes suivantes :

- Extractions des termes.
- Génération de la matrice Article/termes.
- Pondération des descripteurs.

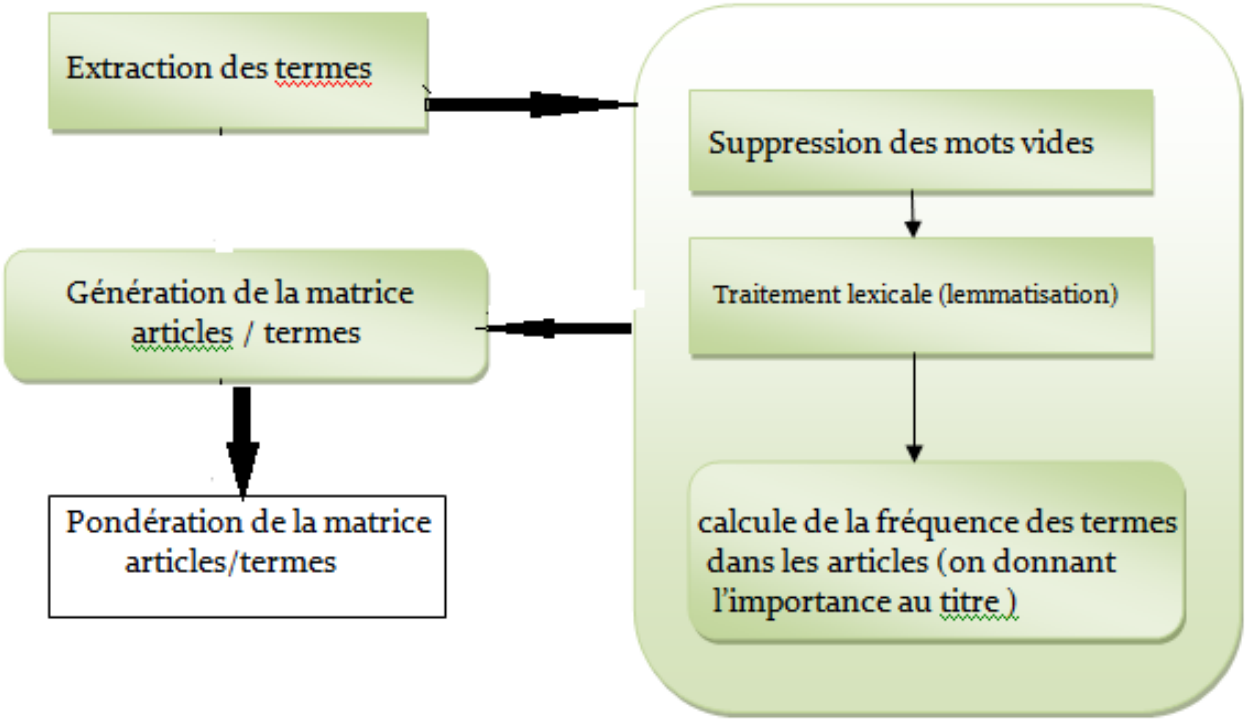

FIGURE 3.5 – Représentation des articles

3.4.2.1 Extraction des termes

Les termes qu'on a choisi a extraire sont ceux des rubrique titre, résumé et mot-clés . On utilisera ces termes pour sélectionner les descripteurs de la matrice de

pondération des articles.

3.4.2.2 Génération de la matrice articles-termes

L'indexation terminologique réduit non seulement l'espace dimensionnel (jusqu'à 50Il vous permet également d'effectuer certaines étapes de prétraitement, telles que la lemmatisation et la suppression des mots vides.

Après avoir extrait le terme du titre, du résumé et des mots-clés de l'article, nous entrons dans la phase de construction de la matrice articles/termes. Pour des raisons d'efficacité, il est important de faire un certain nombre de traitements au lieu de prendre tous les termes détectés :

A. Suppression des mots vides : il existe des termes ou des mots qui ne sont pas importants dans les articles, ces termes peuvent avoir un impact sur notre système ce qui nous oblige de les filtrer.

B. Traitement lexicale (lemmatisation) : processus de regroupement des formes fléchies d'un mot à analyser comme un seul mot un lemme .un lemme est la forme canonique, la forme du dictionnaire ou la forme de citation d'un ensemble de mots.

C. Calcul de la fréquence des termes dans les articles : les termes dans un document n'ont pas la même importance. La pondération permet d'associer des poids en rapport avec l'importance de chaque terme dans la description du contenu du document.

Pour l'exemple précédent on aura ce contenu après l'étape de lemmatisation

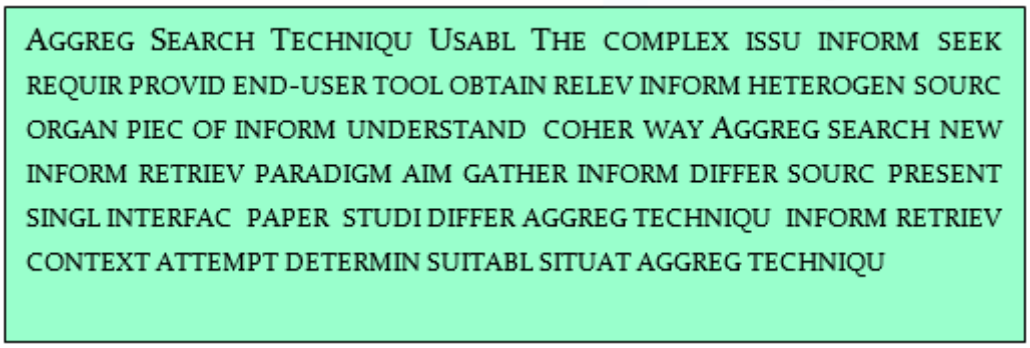

FIGURE 3.6 – Exemple d'un article scientifique aprés lemmatisation

3.4.2.3 Pondération de la matrice

Il existe de nombreuses techniques dédiées à la pondération des termes dans un document .La pondération TF-IDF est de loin la plus utilisée et c'est celle qui donne de meilleurs résultats.

Afin de calculer TF et IDF du terme on utilise les formules suivantes :

Formule TF :

$$TF(T, D) = \frac{NT}{NTm} \tag{3.1}$$

avec NT est la nombre d'apparition du terme dans l'article et NTm lenombre de termes totales dans l'article. Formule IDF :

$$IDF(T, D) = log_{10}\frac{ND}{NDt} \tag{3.2}$$

avec ND est le nombre total des articles et NDt le nombre de document à qui le terme appartient. Formule de poids

$$Poids = TF(T, d) * IDF(T, D) \tag{3.3}$$

On calcule le poids des termes des titres dans les articles. On aura la matrice suivante :

	acycl	automaton	base	calcul	count	cross	cycl	dark	decomposit	determin	...	number	product	prompt	section	singlesourc	sparsitycertifi	stirl	system	tevatron	unlabel
0	0.000000	0.000000	0.000000	0.333333	0.000000	0.333333	0.000000	0.000000	0.00000	0.000000	...	0.000000	0.333333	0.333333	0.333333	0.000000	0.00000	0.000000	0.000000	0.333333	0.000000
1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.57735	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.57735	0.000000	0.000000	0.000000	0.000000
2	0.000000	0.000000	0.333333	0.000000	0.000000	0.000000	0.000000	0.333333	0.00000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.333333	0.000000	0.000000
3	0.333333	0.333333	0.000000	0.000000	0.333333	0.000000	0.333333	0.000000	0.00000	0.333333	...	0.333333	0.000000	0.000000	0.000000	0.333333	0.00000	0.333333	0.000000	0.000000	0.333333
4	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000

5 rows × 32 columns

FIGURE 3.7 – Exemple de pondération de la matrice article-terme

3.4.3 Génération des clusters

Après avoir représenté les articles dans une matrice Terme/article qui représente chaque article par un vecteur de poids correspondant à l'importance de chaque terme du corpus dans l'article, On est prét pour l'étape de Clustering proprement dit.
Il est temps maintenant d'appliquer un Algorithme de Clustering afin de regrouper dans des Clusters.

Cette étape nécessite un certain nombre de traitements. Tout d'abord on doit appliquer la méthode SVD (singular value decomposition) qui permet de réduire l'espace de représentation (réduire la taille de la matrice terme/article) pour améliorer les performances en terme de temps d'exécution du Clustering.

Maintenant que les articles sont représentés dans un format qui peut être interprété par des algorithmes d'apprentissage non supervisé. Ces algorithmes nécessitent une connaissance a priori du nombre optimal du cluster. Dans notre système on a choisi l'algorithme K-means pour la classification des articles cet algorithme prend la mesure de similarité cosinus comme entré, on juge ce choix par plusieurs avantages et critères :

• Algorithme rapide pour le clustering des données.
• Facilité de l'implémentation.
• K-means produit des clusters plus serrés que le clustering hiérarchique.

Afin de trouver le nombre optimal K (nombre cluster)on a choisi la méthode du coude (Elbow Method) L'algorithme prend donc en entrée la matrice de pondération

et le K optimal, K-Means fait plusieurs itérations jusqu'a ce que les centres des clusters s'arrêtent, enfin chaque article sera affecté à son cluster comme le montre la figure ci-dessous :

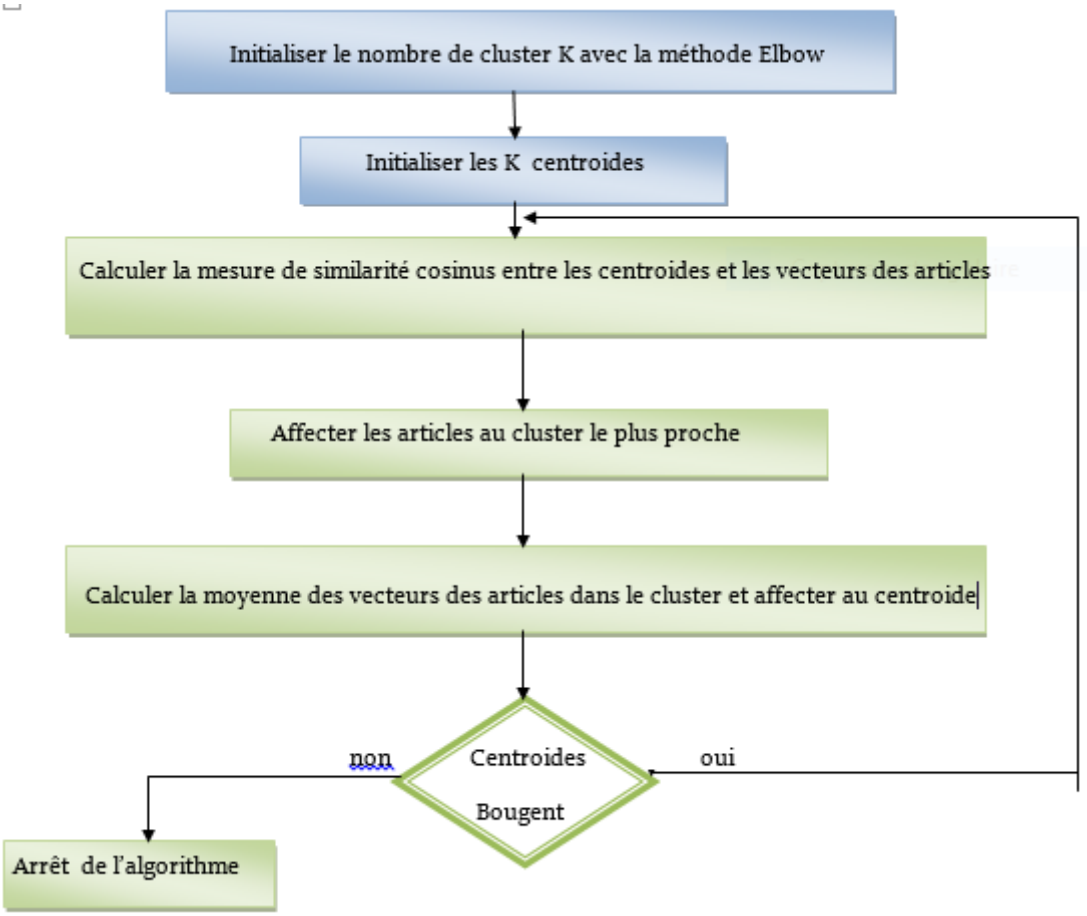

FIGURE 3.8 – Organigramme pour algroithme de clustering

	acycl	automaton	base	calcul	count	cross	cycl	dark	decomposit	determin	...	product	prompt	section	singlesourc	sparsitycertifi	stirl	system	tevatron	unlabel	id_cluster
0	0.000000	0.000000	0.000000	0.333333	0.000000	0.333333	0.000000	0.000000	0.00000	0.000000	...	0.333333	0.333333	0.333333	0.000000	0.00000	0.000000	0.000000	0.333333	0.000000	1
1	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.57735	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.57735	0.000000	0.000000	0.000000	0.000000	2
2	0.000000	0.000000	0.333333	0.000000	0.000000	0.000000	0.000000	0.333333	0.00000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.333333	0.000000	0.000000	1
3	0.333333	0.333333	0.000000	0.000000	0.333333	0.000000	0.333333	0.000000	0.00000	0.333333	...	0.000000	0.000000	0.000000	0.333333	0.00000	0.333333	0.000000	0.000000	0.333333	1
4	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	...	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0

5 rows × 33 columns

FIGURE 3.9 – Exemple de résultat de clustering des articles

A la fin de la classification, nous aurons des articles similaires qui parlent sur les mêmes topics groupées dans les mêmes clusters.

3.4.4 Représentation des résultats

Après avoir générer les différents Clusters, il ne reste que la partie de présentation des résultats à l'utilisateur.

Comme notre but est d'éviter la présentation en listing d'article, la page de résultats doit être organisée en deux parties :

Une partie pour afficher les différents clusters avec pour chaque cluster une description de son contenu. La description du contenu du Cluster doit refléter le contenu de ses articles pour mieux guider l'utilisateur dans l'exploration de l'espace de résultats.

A chaque fois que l'utilisateur choisit un Cluster, les articles de ce cluster doivent etre affichés sur la deuxième partie.

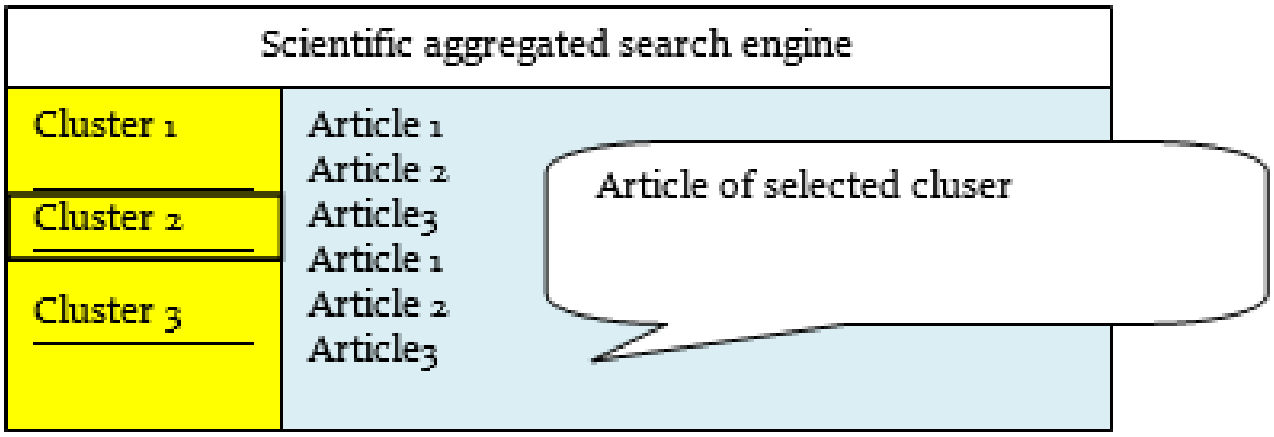

FIGURE 3.10 – Représentation des résultats

Comment générer un titre (une description) pour chaque cluster ?

L'article le plus proche du centroide du cluster est le mieux indiqué pour représenter le contenu du cluster. Chaque Cluster aura comme titre le titre de l'article le plus représentatif du cluster.

3.5 Conclusion

Nous avons présenté dans ce chapitre la technique d'agrégation que nous proposons pour effectuer la classification automatique de document par sujet ,Au début nous

avons défini la classification automatique ainsi que ses objectifs puis nous avons exposé le processus de classification automatique de document.

Ensuite nous avons détaillé notre architecture générale du système en commencant par la collection et nettoyage des articles , et finissent par la représentation des résultat.

Dans le chapitre suivant, on va présenter notre implémentation et évaluation du système.

Chapitre 4

Implémentation et évaluation

4.1 Introduction

Après la présentation de la solution et les choix de conception dans le chapitre précédent, nous allons à présent présenter l'implémentation de notre système en termes d'environnement de développement, d'outils et de langage utilisé ainsi qu'en terme d'évaluation des résultats obtenu, notamment sur les performance et la qualité du Clustering des résultats de recherche.

4.2 Environnement de développement

4.2.1 Bibliothèques utilisées

Nous avons utilisé plusieurs bibliothèques telles que :

Scikit-learn

Scikit-learn est un module Python intégrant une large gamme d'algorithmes d'apprentissage automatique de pointe pour les problèmes supervisés et non supervisés à moyenne échelle. Ce package vise à apporter l'apprentissage automatique aux non-spécialistes à l'aide d'un langage de haut niveau à usage général. L'accent est mis sur la facilité d'utilisation, les performances, la documentation et la cohérence de l'API. Il a des dépendances minimales et est distribué sous la licence BSD

simplifiée, encourageant son utilisation dans des contextes académiques et commerciaux.
scikit-learn est construit sur les bibliothèques python populaires Numpy et SciPy .Numpy étend python pour prendre en charge des opérations efficaces sur de grands tableaux et des matrices multidimensionnelles.Scipy fournit des modules pour le calcul scientifique. La bibliothèque de visualisation matplotlib est souvent utilisée conjointement avec scikit-learn[17].

Pandas

pandas est un package Python fournissant des structures de données rapides, flexibles et expressives conçues pour faciliter le travail avec des données « relationnelles » ou « labellisées » à la fois simples et intuitives. Il vise à être le bloc de construction fondamental de haut niveau pour faire analyse pratique de données du monde réel en Python.[18]

NLTK

NLTK est le Natural LanguageToolkit, une bibliothèque Python complète pour le langage naturel traitement et analyse de texte. Conçu à l'origine pour l'enseignement, il a été adopté dans le l'industrie pour la recherche et le développement en raison de son utilité et de l'étendue de sa couvertur.[19]

Matplotlib

Est une bibliothèque permettant de créer des tracés 2D de tableaux en python .Bien qu'il ait ses origines dans l'émulation des commandes graphiques, il est indépendant de Matlab et peut être utilisé de manière pythonic.orientée objet. Bien que Matplotlib soit écrit principalement en python pur, il fait un usage intensif de Numpy et d'autres codes d'extension pour fournir de bonnes performances même pour les

grandes tableaux.

Maplotlib est conçu avec la philosophie que vous devriez être capable de créer tracés simples aves seulement quelques commandes, ou une seule, si vous voulez voir un histogramme de vos données, vous ne devriez pas avoir besoin d'instancier des objets, d'appeler des méthodes, de définir des propriétés. Ça devrait juste marcher.[20]

4.2.2 Langage du développement

Nous utilisons le langage de programmation python pour implémenter notre solution. Le langage de programmation Python a été créé en 1989 par Guido van Rossum, aux Pays-Bas.La première version publique de ce langage a été publiée en 1991. La dernière version de Python est la version 3. Plus précisément, la version 3.7 a été publiée en juin 2018. cessera d'être maintenue après le 1er janvier 2020.. La Python Software Foundation 1 est l'association qui organise le développement de Python et anime la communauté de développeurs et d'utilisateurs. Ce langage de programmation présente de nombreuses caractéristiques intéressantes :

— Il est multiplateforme. C'est-à-dire qu'il fonctionne sur de nombreux systèmes d'exploitation : Windows, Mac OS X, Linux, Android, iOS, depuis les mini-ordinateurs Raspberry Pi jusqu'aux supercalculateurs. — Il est gratuit. Vous pouvez l'installer sur autant d'ordinateurs que vous voulez (même sur votre téléphone !).

— C'est un langage de haut niveau. Il demande relativement peu de connaissance sur le fonctionnement d'un ordinateur pour être utilisé. — C'est un langage interprété. Un script Python n'a pas besoin d'être compilé pour être exécuté, contrairement à des langages comme le C ou le C++.

— Il est orienté objet. C'est-à-dire qu'il est possible de concevoir en Python des entités qui miment celles du monde réel (une cellule, une protéine, un atome, etc.) avec un certain nombre de règles de fonctionnement et d'interactions._ — Enfin, il est très utilisé en bioinformatique et plus généralement en analyse de données. Toutes ces caractéristiques font que Python est désormais enseigné dans de nombreuses formations, depuis l'enseignement secondaire jusqu'à l'enseignement supérieur.[21]

4.2.3 Plateforme et environnement de développement

Google Colab

Colaboratory, souvent raccourci en « colab », est un produit de google Research.Colab permet à n'importe qui d'écrire et d'exécuter le code python de son choix par le biais du navigateur. C'est un environnement particulièrement adapté au machine Learning, à l'analyse de données et à l'éducation .En termes plus techniques, colab est un service héberge de notebooks jupyter qui ne nécessite aucune configuration et permet d'accéder sans frais à des ressources informatiques, dont des GPU [22].parmi les caractéristiques de GPU on trouve que :

-Les GPU sont très rapides et performants, car ils sont efficaces pour la multiplication de matrices et la convolution. En guise d'explication, on cite souvent le parallélisme. Ce n'est toutefois pas la seule raison.

- GPU est qu'il est optimisé en bande passant.

- GPU est facilement programmable. Pour toutes ces raisons, les processeurs graphiques sont idéaux pour le Deep Learning et les traitements Big Data.

4.3 Evaluation des résultats

Après avoir effectué plusieurs applications pour pouvoir faire des évaluations précise, nous allons voir les résultats de ses derniers. Comme on la mentionnée dans le chapitre précédent sur la technique d'agrégation appliquer qui suit une procédure très simple pour classer un ensemble de données à travers un certain nombre(K) de cluster.il existe une méthode pour pouvoir trouver le meilleur K, c'est la méthode du coude (Elbow method.)

4.3.1 Comment trouver le meilleur K

Pour utiliser l'algorithme de k-means, les utilisateurs sont censés de trouver le meilleur K (nombre de groupes) afin d'acquérir des bons résultats. Cependant, il existe une méthode pour trouver le meilleur K, c'est la méthode du coude (Elbow method).

L'idée de cette technique, est d'exécuter l'algorithme K-means sur une plage de valeur de K et calculer le SSE (somme carré de la distance entre un document du cluster et son centroide).

Après, on trace le graphe (qui correspond à un bras), la valeur qui se situe dans le coude du bras c'est la valeur du meilleur K de l'algorithme[23].

La figure ci-dessous montre le résultat du K dans notre cas :

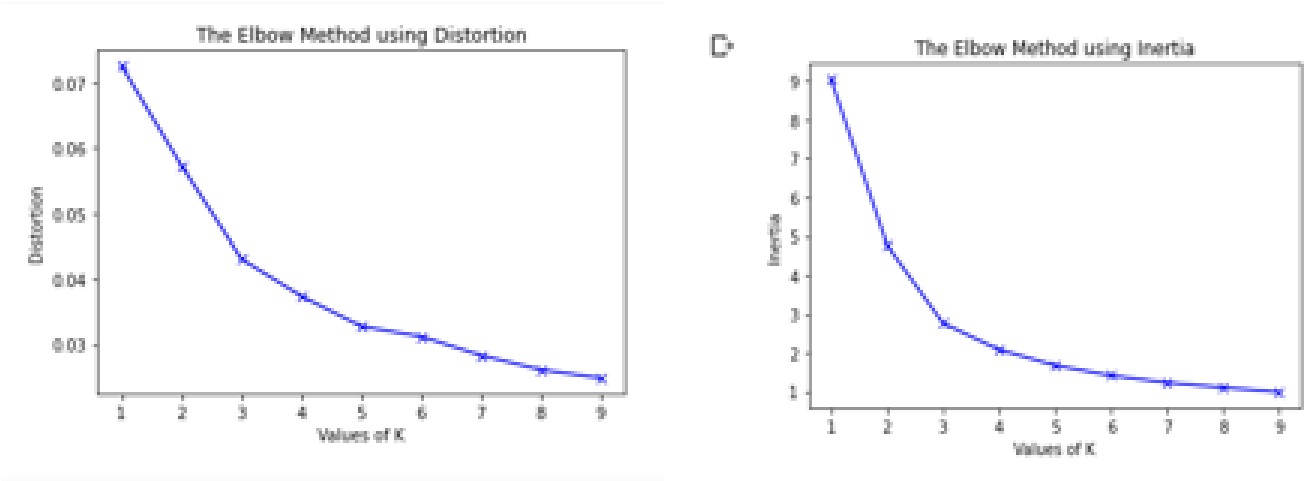

FIGURE 4.1 – Le nombre de cluster K

Ensuite après avoir calculer le nombre de cluster K , on appliquer l'algorithme sur notre dataset, dans notre cas on appliquer le K-means sur 1000 articles . La figure ci-dessous montre le résultat de l'algorithme K-means :

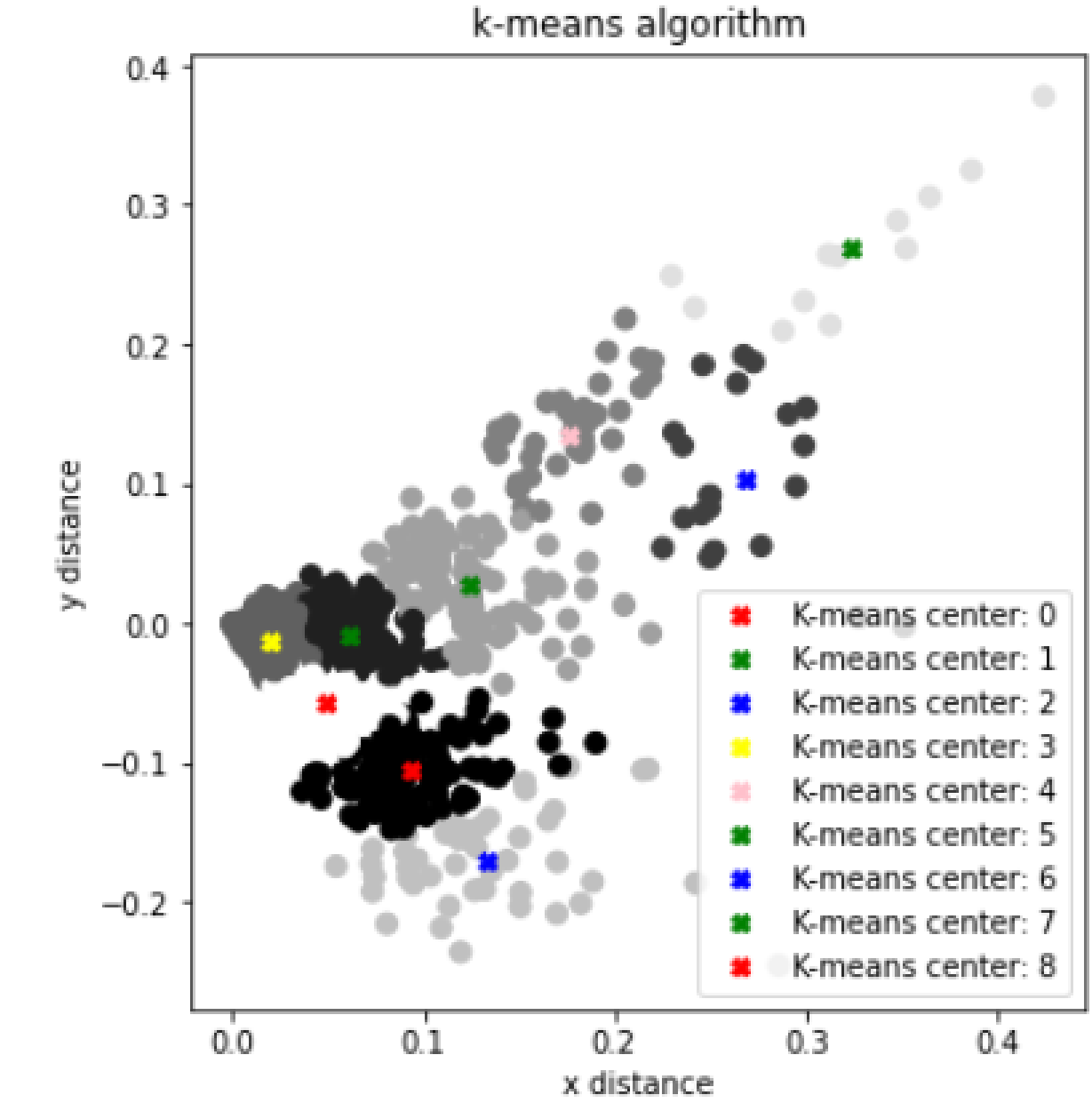

FIGURE 4.2 – Le nombre de cluster K

4.3.2 Coefficient de silhouette

le **coefficient de silhouette** est une mesure de qualité d'une partition d'un ensemble de données en classification automatique 1.Pour chaque point, son coefficient de silhouette est la différence entre la distance moyenne avec les points du même groupe que lui (cohésion) et la distance moyenne avec les points des autres groupes voisins (séparation). Si cette différence est négative, le point est en moyenne plus proche du groupe voisin que du sien : il est donc mal classé.

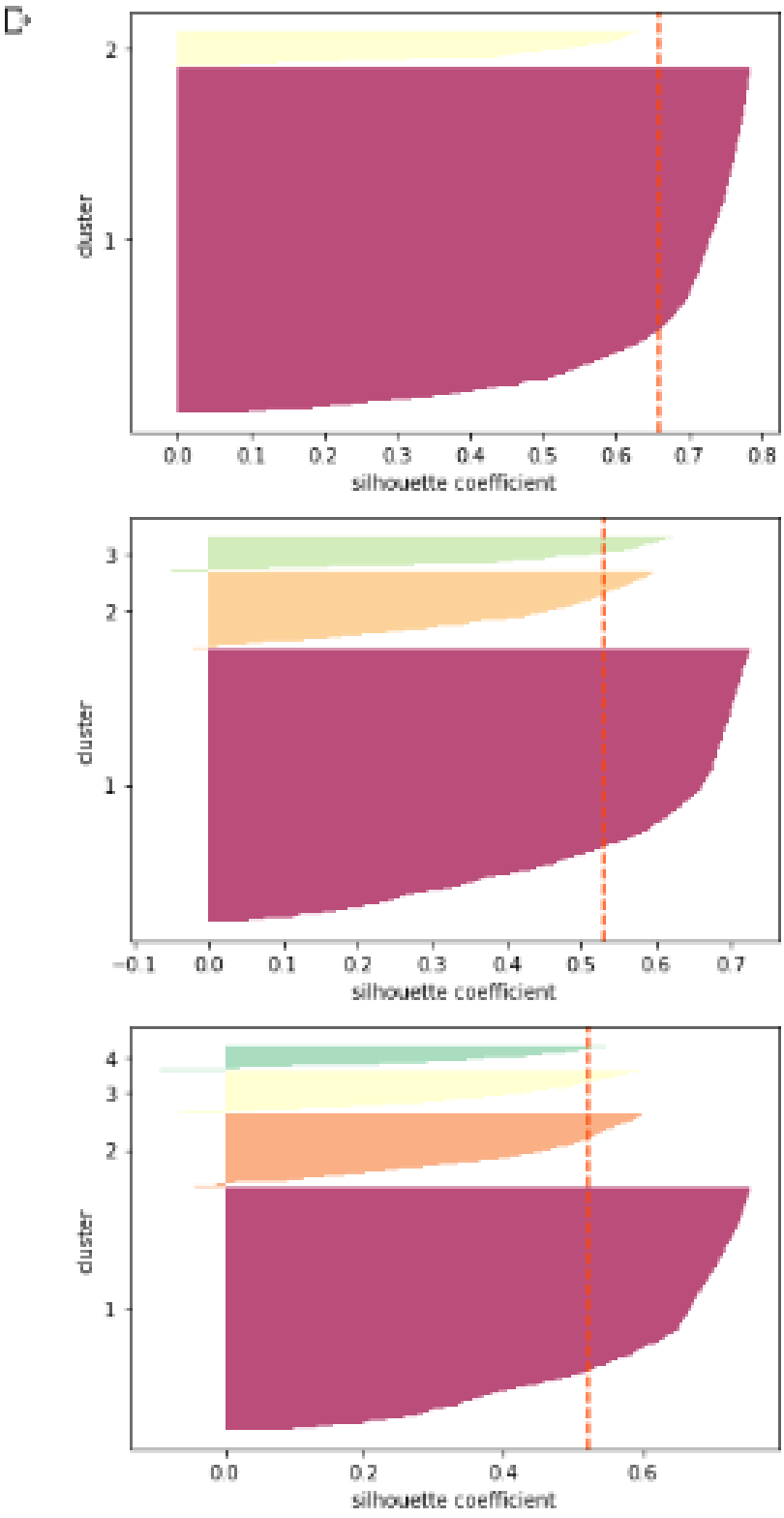

FIGURE 4.3 – Résultat de coefficient silhouette

Les diagrammes de silhouette sont tracés pour le nombre de clusters spécifiés comme 2, 3 et 4, respectivement. La ligne pointillée rouge représente le coefficient de silhouette moyen. Si les grappes sont correctement séparées, "l'épaisseur" des silhouettes dans chaque grappe a tendance à être proche de l'égalité. Dans la figure ci-dessus, "l'épaisseur" de la silhouette est paire lorsque le nombre de clusters est de 3, et la valeur moyenne du coefficient de silhouette est la plus élevée. De cela, nous pouvons conclure que le nombre optimal de clusters est de 3.

4.4 Conclusion

Nous avons consacré ce chapitre pour l'implémentation de notre système. Ensuite nous avons défini l'environnement et l'ensemble d'outils utilisés.. Et enfin Nous avons fait une évaluation et résultats de notre algorithme.

Conclusion et Pererspectives

Le travail décrit dans cet article s'inscrit dans le contexte général de la recherche d'information (RI), en particulier dans le cadre de la RI agrégée. Aggregated search est un nouveau paradigme de recherche d'informations qui, contrairement à l'RI classique, ne se limite pas à renvoyer une liste de documents pertinents à l'utilisateur, mais peut également générer automatiquement un espace de réponse agrégé cohérent et bien organisé. Sélectionner et agréger automatiquement les unités d'information pertinentes (nuggets) recherchées dans plusieurs sources d'information pour produire des réponses complètes, et bien organiser pour les utilisateurs finaux .

Nous avons dans ce modeste travail concevoi et développer un système de recherche d'information agrégée pour la recherche d'articles scientifique en utilisant une technique d'agrégation de document, Comme le nombre et le volume des documents numériques s'accroissant de façon exponentielle, on a besoin de les catégoriser afin de facilité leur manipulation. La classification automatique de textes a pour objectif de regrouper les textes similaires c'est à dire thématiquement proches, au sein d'un même ensemble. En d'autres termes, trouver un algorithme permettant d'assigner un texte à une classe avec le plus grand taux de réussite possible.

Pendant l'implémentation de notre application nous avons rencontré un certain nombre de difficultés, notamment la celle liées aux API des sources d'information utilisées qui étaient payantes pour la plupart. Qui nous a obligé de télécharger un dataset et travailler avec .

Comme perspectives, à court terme nous souhaitons d'implémenter notre application web par les fonctionnalités qu'on a besoin .

Bibliographie

[1] Krichen,Ines and Kopliku, Arlind and Pinel, Sauvagnat, Karen and Boughanem, Mohand. "Une approche de recherche d'attributs pertinents pour l'agrégation d'information",Document numerique,Lavoiser,2012

[2] http ://halice au pays de merveiles.com(consulté le 02/03/2022)

[3] Kopliku, Arlind and Pinel-Sauvagnat, Karen and Boughanem, Mohand. "Aggregated search : A new information retrieval paradigm",ACM Computing Surveys (CSUR),ACM New York, NY, USA,2014

[4] Pinel-Sauvagnat, Karen."De la recherche de granules documentaires à l'agrégation d'information",Université Paul Sabatier (Toulouse 3),2018.

[5] Chellali Tarek,Haddouche Rabah."Systéme de Génération Automatique de revue de presse",Université des Sciences et de Technologie Houari Boumedine",2015.

[6] Aggoun Meriem,Bensidi Rym."Génération automatique des brochures touristiques",Université des Sciences et de Technologie Houari Boumedine,2015.

[7] https ://www.researchgate.net/publication/260311480(consulté le 02/07/2022).

[8] Melero, Eduardo."Are workplaces with many women in management run differently ?",Journal of Business Research,Elsevier,64,4,385–393,2011.

[9] M BEN ROMDHANE."Analyse des publications scientifiques : caractéristiques", 1995-1996

[10] Heliment Ahlem."Analyse morphosemantique de la description dans les ecrits scientifique",Université Mohamed Khider de Biskra,2021.

[11] Smail lamia."La figure de l'auteur entre la subjectivité énonciative et l'objectivité discursive
dans les articles scientifiques",Université kasdi merbah ouargla,2014.

[12] Boudjelal,Meftah."Rédiger et publier un article scientifique : techniques et enjeux",2014.

[13] Beaudry,éric."Rédiger et publier un article scientifique",2011.

[14] https ://htpratique.com/publications-academiques/(consulté le 03/09/2022).

[15] Toucherifte, Samira."Etude comparative en classification non-supérvisée",2011

[16] Korde Vandana. "Text Classification and Classifiers :A Survey." International Journal of Artificial Intelligence Applications. 3. Pages 85-99, 2012.

[17] Hackeling,Gavin."Mastering Machine Learning with scikit-learn",Packt Publishing Ltd,2017.

[18] McKinney, Wes and Team, PD."Pandas-Powerful python data analysis toolkit",Pandas—Powerful Python Data Analysis Toolkit,1625,2015.

[19] Perkins, Jacob."Python text processing with NLTK 2.0 cookbook",PACKT publishing,2010.

[20] Barrett, Paul and Hunter, John and Miller, J Todd and Hsu, J-C and Greenfield, Perry."matplotlib–A Portable Python Plotting Package",Astronomical data analysis software and systems XIV,347,91,2005.

[21] https ://hal.archives ouvertes.fr/hal03264103v1/document(consulté le 02/09/2022).

[22] https ://research.google.com/colaboratory/faq.html hl=fr(consulté le 02/09/2022).

[23] Gildas Tagny Ngompe. "Méthodes D'Analyse Sémantique De Corpus De Décisions Jurisprudentielles", Mines Alès Ecole Mines - Télécom, These doctorat, 2020.