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** : Génie des systèmes informatique*

Thème

Modélisation des profils thématique dans un réseau social académique

---

Proposé et encadré par :

- Mr Boussaadi Smail

Réalisé par :

- Houssou Sabrina
- Ouazani Kamilia

2019/2020

Remerciement

Au tout puissant ALLAH qui nous a donné la force, le courage et la santé afin de réaliser ce modeste travail.

Nous tenons à remercier en premier lieu notre chère professeur et directeur de mémoire Ms BOUSSAADI pour le temps qu'il nous a consacré, pour ses précieux conseils, sa patience, sa disponibilité et pour sa bienveillance. Veuillez trouver monsieur l'expression de notre sincère gratitude.

Nous remercions également les membres du jury pour avoir jugé et noté notre modeste travail

A tous ceux qui ont contribué de près ou de loin à la réalisation de ce travail.

Dédicaces

C’est avec un très grand bonheur et une immense joie que nous dédions ce modeste travail a nos chers parents, qui ont été présent pour nous et qui nous ont épaulé tout au long de nos vie, nous leur devons ce que nous sommes aujourd’hui.

.

A nos fréres et soeurs qui nous ont tant encouragé

A nos amis qui nous ont toujours soutenu

Résumé

les réseaux sociaux académiques sont devenus des plateformes incontournables pour tout scientifique. Leur popularité est due entre autres à certaines fonctionnalités que les scientifiques jugent nécessaires et intéressantes comme une plus large visibilité numérique, la possibilité de collaborer et de communiquer avec des pairs qui partagent les mêmes domaines d'intérêt.

Dans ce présent travail, nous nous intéressons à l'identification des domaines d'intérêt (ou d'expertise) des scientifiques membres sur ces plateformes. Afin de modéliser chaque chercheur par son profil thématique. Pour faire on s'appuie sur un outil d'extraction des structures sémantiques latentes dans un corpus, une technique du domaine de machine learning non supervisé qui est designée par la modélisation de sujets (topic modeling).

Mots clés : Réseaux sociaux , Plateformes sociales académiques , modélisation thématique , allocation de dirichlet latent, profil utilisateur

Abstract

academic social networks have become essential platforms for any scientist. Their popularity is partly due to certain features that scientists find necessary and crucial, such as greater digital visibility, the ability to collaborate and communicate with peers who share the same areas of interest.
In this present research work, we are aim at identifying areas of interest (or expertise) of scientific members on these platforms. This is achived in order to model each researcher by their thematic profile. Moreover, in order to reach these goals, we based our work on a tool for extracting latent semantic structures in a corpus, this tool is a technique in the field of non-supervisory machine learning which is designated by topic modeling.

Keywords : Social networks, Academic social platforms, thopic modeling, latent dirichlet allocation, user profile

Table des matiéres

Table des matiéres ..... **ii**
Table des figures ..... **iv**
Liste des tableaux ..... **v**
Liste des abréviations ..... **vi**
Introduction générale **1**
1.Généralités sur les réseaux sociaux **3**
1.1 Introduction 3
1.2 Réseaux sociaux numériques 3
1.3 Typologie de réseaux sociaux 4
1 . 3 . 1 Les réseaux sociaux Généralistes 4
1.3.2 Les réseaux sociaux spécialisés 4
1.4 Pourquoi s'intéresse-t-on à l'analyse des RSNs ? 5
1.5 Réseaux sociaux académiques 6
1.5.1 Exemple de plateformes académiques 6
1.5.3 Topologie des RSAs 9
1.5.3.1 RSAs Homogènes 9
1.5.3.2 RSAs Hétérogènes 10
1.6 Conclusion 11
2.Etat de l'art **12**
2.1 Introduction 12
2.1 Représentation vectorielle des textes 12
2.2 Modélisation thématique 13
2.2.1 LSA: L'analyse sémantique latente 13
2.2.2 PLSA: L'analyse sémantique latente probabiliste 14
2.2.2.1 Définition 14
2.2.2.2 Principe du PLSA 14
2.2.3 LDA: allocation de Dirichlet latent 15
2.2.3.1 Définition 15
2.2.3.2 Principe du LDA 16
2.3 Profil utilisateur 18
2.3.1 Profilage 19
2.3.2 Collecte de Données 19
2.3.2.1 Méthode explicite 20
2.3.2.2 Méthode Implicite 20
2.3.2.3 Prétraitement de données 21
2.3.2.3.1 Techniques de prétraitement de données 21
2.3.3 Construction du modèle utilisateur 24
2.3.3.1 Mots clés 24

2.3.3.2 Réseau sémantique 25
2.3.3.3 Profil conceptuel 25
2.3.4 Mise à jour du modèle utilisateur 25
2.3.5 Structure d'un article scientifique 26
2.4 conclusion 27
3.Approche proposée 28
Partie: I 28
3. I.1. Introduction 28
3. I.2 Description générale de l'approche proposée 28
3. I.2 .1 Etape 1: collecte de données 29
3. I.2.2 Etape 2: prétraitement 30
3. I.2.3 Etape 3: l'entraînement du modèle LDA 31
3. I.3 Scénario de modélisation 32
3. I.3.1 Cas 1: Chercheur junior 32
3. I.3.2 Cas 2: Chercheur senior 35
Partie II 37
3.II.1 Introduction 37
3.II.2 Environnements et technologies 38
3.II.2.1 Environnements de développement 38
3.II.2.1.1 Jupyter 38
3.II.2.1.2 python 38
3.II.2.1.3 Canva 38
3.II.2.2 Outils utilisés 38
3.II.3 Implémentation (code + Execution) 39
3.II.4 Discussion des Résultats 47
3.II.5 Exemple illustratif 47
3.3 Conclusion 48
Conclusion générale 49
Perspectives 49
Bibliographie 50

Table des figures

FIGURE 1-LES CATÉGORIES DE RÉSEAUX SOCIAUX NUMÉRIQUES (RSNS) 5
FIGURE 2-CO-AUTEUR 9
FIGURE 3-CO-CITATION 10
FIGURE 4-COUPLAGE BIBLIOGRAPHIQUE 11
FIGURE 5-REPRÉSENTATION VECTORIELLE DE DOCUMENT 13
FIGURE 6-GRAPHE DE FONCTIONNEMENT DE PLSA 15
FIGURE 7-LDA ENTRÉES / SORTIES 15
FIGURE 8-GRAPHE DE FONCTIONNEMENT DE LDA 17
FIGURE 9-AVANTAGES ET INCONVENIENTS DES MODELES THEMATIQUES 17
FIGURE 10-LES PHASES DU PROFILAGE 19
FIGURE 11-STRUCTURE D'UN ARTICLE SCIENTIFIQUE 27
FIGURE 12-ETAPES DE L'APPROCHE PROPOSEE 29
FIGURE 13-COLLECTE DE DONNÉES (CRAWLING) 30
FIGURE 14-PRÉTRAITEMENT ET VECTORISATION 30
FIGURE 15-CRÉATION DE LA MATRICE CHERCHEUR-TOPIC 31
FIGURE 16-ETUDE DES CAS 32
FIGURE 17-MODÉLISATION DES CHERCHEURS JUNIORS 33
FIGURE 18-LES QUATRE MODÈLES LDA 34
FIGURE 19-ETUDE COMPARATIVE / CAS 1 35
FIGURE 20-MODÉLISATION DES CHERCHEURS SENIORS 36
FIGURE 21-ETUDE COMPARATIVE / CAS 2 37
FIGURE 22-LES DIFFERENTES BIBLIOTHÈQUES 39
FIGURE 23-PARTIE DU DATASET DES CHERCHEURS JUNIORS 39
FIGURE 24-PARTIE DU DATASET DES CHERCHEURS SENIORS 40
FIGURE 25-RÉSULTAT APRÉS LE PRÉTRAITEMENT DE DONNÉES 40
FIGURE 26-CHEMIN VERS LE MALLET 40
FIGURE 27-PREPARATION DES ENTREES DE LDA 41
FIGURE 28- ENTRAINEMENT DU MODELE LDA 41
FIGURE 29-PARTIE DE MATRICE CHERCHEUR-TOPIC ISSUE DU DEUXIÈME MODÈLE LDA /CAS1 44
FIGURE 30-PARTIE DE LA MATRICE DOCUEMNT-TOPIC ISSUE DU DEUXIÈME MODÈLE LDA / CAS2 44
FIGURE 31-PARTIE DE LA MATRICE CHERCHEUR-TOPIC ISSUE DE L'AGRÉGATION PAR MAXIMUM 45
FIGURE 32-PARTIE DE LA MATRICE CHERCHEUR-TOPIC ISSUE DE L'AGRÉGATION PAR MOYENNE 45
FIGURE 33-RESUME DE L'APPROCHE PROPOSEE 46
FIGURE 34-MODELISATION D'UN CHERCHEUR SENIOR 47
FIGURE 35-MODELISATION D'UN CHERCHEUR JUNIOR 48

Liste des tableaux

TABLEAU 1-ACADEMAI VS RESEARCHGATE [6] [7] [8] [13] 8

TABLEAU 2-CONTENUE D'UN MODÈLE UTILISATEUR EN GÉNÉRAL ET UN MODÈLE D'UN CHERCHEUR 18

TABLEAU 3-AVANTAGES ET INCONVÉNIENTS DES MÉTHODES DE COLLECTE DE DONNÉES 21

TABLEAU 4-EXEMPLES DE QUELQUES TECHNIQUES DE PRÉTRAITEMENT 23

TABLEAU 5-RÔLE DES TECHNIQUES DE PRÉTRAITEMENT DES DONNÉE 24

TABLEAU 6-RÉSULTAT OBTENUE DANS LE PREMIER CAS 34

TABLEAU 7-RÉSULTATS OBTENUE DANS LE DEUXIÈME CAS 36

TABLEAU 8-LES TOPIC EXTRAITS AVEC LE DEUXIÈME MODÈLE LDA SUR LE DATASET DES CHERCHEURS JUNIORS 42

TABLEAU 9-LES TOPIC EXTRAITS AVEC LE DEUXIÈME MODÈLE LDA SUR LE DATASET DES CHERCHEURS SENIORS 43

.

Liste des abréviations

RSNs	Les Réseaux Sociaux Numériques
RSAs	Les Réseaux Sociaux Académique
SNA	Social Network Analytic
VSM	Vector Space Modèle
TF-IDF	Term Frequency-Inverse Document Frequency
SVD	Décomposition en Valeurs Singulières
LSA	Latent Semantic Analysis.
PLSI	Probabilist Latent Symantic Analysis.
LDA	Latent Dirichlet Allocation.
POS	Part Of Speech
TAL	Traitement Automatique du Laguaage naturel

Introduction générale

Contexte du travail

De nos jours, les technologies de l'information, de la communication et le web social et collaboratif en particulier, ont révolutionné notre façon de vivre, de penser et d'agir. À travers les réseaux sociaux numériques (RSNs) qui sont désormais omniprésents dans la vie quotidienne de chacun de nous. Ces réseaux sont dérivés en plusieurs plateformes, dans la mesure où ce présent travail est basé sur les réseaux sociaux académiques (RSAs).

L'utilisation des RSAs a vu une croissance exponentielle ces dernières années, ce qui a poussé en effet les chercheurs à s'intéresser à l'exploitation et l'analyse des données contenant dans les plateformes de ces réseaux. Dans ce type d'étude, chaque utilisateur doit être représenté efficacement par un profil qui reflète son axe de recherche ou domaine d'intérêt.

Un profil thématique de chercheur est construit à partir de son interaction sur son RSA. Pour cela, on s'appuie sur des informations issus de son activité (historique) comme les articles publiés, téléchargés… Dans ce mémoire, nous nous focalisons sur l'extraction des thèmes pour pouvoir modéliser le profil thématique d'un scientifique.

Problématique

Dans un RSA, le profil thématique d'un chercheur est défini entre autres par les sujets de recherche qui sont véhiculés dans les articles qui l'intéressent. Néanmoins, la majorité des chercheurs ne fournissent pas explicitement des informations sur leur domaine d'expertise et dans certains cas, ces domaines d'intérêt évoluent ou changent avec le temps, ce qui engendre un problème dans l'identification réelle et précise des sujets sur lesquels un chercheur intervient, cela nous permet d'évoquer la problématique suivante:

Comment modéliser un profil thématique pertinent qui reflète efficacement le domaine d'intérêt d'un chercheur dans le contexte des RSAs ?

Objectifs visés

À travers ce mémoire, nous nous sommes fixé comme objectif le développement d'une nouvelle approche de modélisation des profils thématiques des chercheurs dans un réseau social académique. Cette approche est focalisée principalement sur l'extraction des sujets latents (intérêts) qui intéressent les chercheurs et surtout de calculer leurs poids par rapport aux autres intérêts de l'ensemble. Ainsi, le but de notre recherche est de construire les profils thématiques les plus précis en représentant chaque chercheur par un vecteur d'intérêts dont les composants traduisent, avec précision l'importance que porte chaque thème pour ce chercheur.

Structure générale du mémoire

Nous avons structuré notre travail en trois parties essentielles:

La première partie concerne le chapitre n°1 intitulé de « généralités sur les réseaux sociaux numériques », comme son nom l'indique celui-ci est consacré à des notions de base sur les RSNs, ce qui comprend les diverses catégories des RSNs ainsi que l'étude et l'analyse de ces réseaux. Nous aborderons par la suite un type particulier des RSNs qui est les RSAs en mentionnant quelques exemples, suivi des différentes topologies des RSAs.

La deuxième partie est consacrée au chapitre n°2, celui-ci se concentre sur notre état de l'art qui est basé principalement sur la modélisation des profils thématiques des chercheurs. Nous commençons d'abord par donner une brève définition de la représentation vectorielle de textes ainsi que la modélisation thématique, nous élaborons ensuite les divers aspects des modèles thématiques. Enfin, nous clôturons ce chapitre par les notions fondamentales relatives au profil utilisateur, son contenu, notamment les différentes phases du profilage.

La troisième partie est destinée au chapitre n°3 intitulé de « Approche et contribution », ce dernier est divisé en deux sous-parties principales, nous présentons dans la première, l'approche que nous avons proposé pour modéliser thématiquement les profils des chercheurs. Nous décrivons dans un premier temps de manière générale notre approche, puis nous détaillons chaque étape du processus appliqué. En outre dans la deuxième partie, nous présentons les résultats obtenus suivi des outils utilisés ainsi que nos perspectives.

Chapitre 1

Généralités sur les réseaux sociaux

1.1 Introduction

L'apparition du phénomène Web social et collaboratif dans les années 2000 a changé le monde. Aujourd'hui plus de 60 % de la population mondiale utilise l'internet pour de nombreuse raisons tels l'achat en ligne, la consultation des comptes bancaires, la réservation des billets d'avion, notamment pour bénéficier des services de réseautage sociales[1]. Avec 3.8 milliards correspondant à 49 % de la population mondiale en janvier 2020. Les réseaux sociaux numériques (RSNs) ont envahi le web et sont de plus en plus populaires, en mettant en évidence leurs nombreux avantages pour la communication et l'accès à l'information. Ils représentent clairement aujourd'hui l'outil privilégié des utilisateurs dans leur quotidien. [15] [16] [2]

Ce chapitre est consacré à des généralités sur les RSNs. Nous commencerons dans un premier temps par les définir en mentionnant leurs différentes catégories, dans un deuxième lieu nous entamerons le concept d'analyse de ces réseaux, notamment son intérêt, nous définisrons par la suite les réseaux sociaux à caractère académique en donnant quelques exemples et enfin, nous parlerons sur la topologie des réseaux sociaux académiques (RSAs).

1.2 Réseaux sociaux numériques

Danah M. Boyd et Nicole B. Ellison identifient les (RSNs) comme des services Web qui permettent aux internautes de construire une identité numérique (profil), définir une liste d'autre

1. Ensemble de moyens virtuel reliant des personnes entre elles

s utilisateurs avec lesquels ils partagent une connexion, aussi afficher et parcourir la liste de leur

Connexions et celles effectuées par d'autres au sein du système. [6]

Du point de vu des analystes, Un RSN correspond à des collections d'individus ou d'organisation et de relation sociale entre eux, il s'agit en effet d'un ensemble de nœuds et d'arêtes. Les relations sociales (arêtes) présentent une grande diversité, les différents types de réseaux ont différents types de relations en fonction de leur degré de force, de leur sens d'orientation... etc. Cette représentation graphique a été conçue principalement dans le but de faciliter l'analyse et l'exploit des données sociales nommé "social network analytic" (SNA) [9].

1.3 Typologie de réseaux sociaux

Il existe plusieurs plateformes de réseaux sociaux. Ces derniers Peuvent être divisés en deux grandes catégories selon les différents services qu'ils proposent : (figure 1)

1.3.1 Les réseaux sociaux Généralistes

Ce type de réseau vise le grand public et permet de partager tout type de contenu (texte, vidéo, image, etc.), de tenir une discussion, et de rester en contact avec les amis, la famille et les collègues.

1.3.2 Les réseaux sociaux spécialisés

Ils ont des finalités plus professionnelles que les réseaux sociaux généralistes. Ils servent à cibler les internautes de manière précise. De ce fait les regroupés selon un thème spécifique tel un métier (vox-avocats pour les avocats), un secteur d'activité (ex talent pharmacie pour la santé), etc. Nous citons dans ce qui suit les types les plus fréquents de réseaux sociaux spécialisés :

- Professionnel : les RSNs à caractère Professionnel permettent le Réseautage qui consiste à former un réseau de contact professionnel avec des gens ayant les mêmes intérêts professionnels. Et mettent en évidence les études et l'expérience professionnelle antérieure, ce qui ressemble à un CV. Ils permettent également la recherche des informations sur des entreprises.
- Partage de contenu : ce type de réseau permet de diffuser l'information sous différents formats: vidéo, visuel, audio …etc.

- Académique (RSAs) : ces réseaux ont été conçus principalement pour répondre aux besoins des scientifiques et institutions de recherche, ils sont destinés à faciliter et à favoriser la communication entre chercheurs. Nous présentons plus de détails sur cette typologie de réseaux dans ce qui suit.

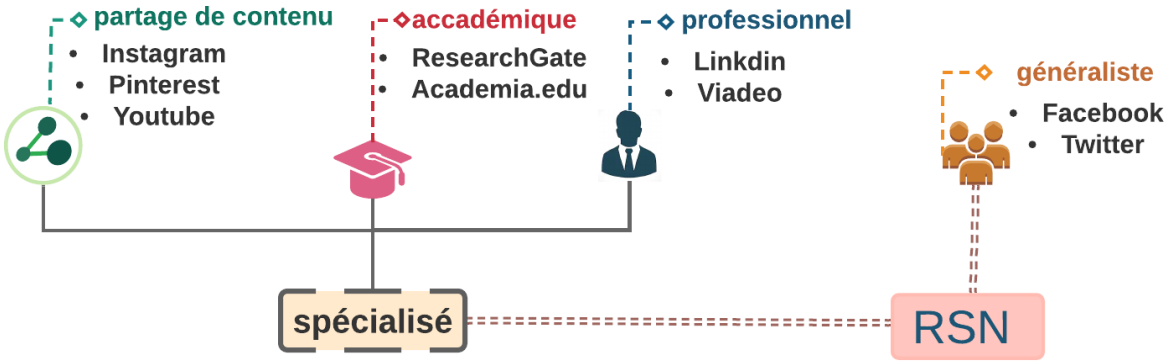

Figure 1-Les catégories de réseaux sociaux numériques (RSNs)

1.4 Pourquoi s'intéresse-t-on à l'analyse des RSNs ?

Les RSNs attirent de plus en plus le grand public, entrainant un défi majeur pour leurs études et analyses qui désignent une technologie d'exploration et de visualisation des données sociales à travers l'identification et la modélisation des interactions entre différentes entités ainsi que la forme et les implications de ces interactions formées au cours du processus de diffusion de l'information dans le réseau [3][4].

Les techniques de SNA sont devenues désormais, l'un des domaines de recherches les plus développés. Elles sont adoptées par les analystes dans le but d'améliorer la sélection des utilisateurs similaires dans les systèmes de recommandation, aider les entreprises à mieux comprendre leurs publics et aiguiller leurs décisions…etc.

SNA devient de plus en plus complexe, d'une façon qu'elle dépasse les capacités des outils informatiques classiques, cela est due à la nature des données sociales qui ont atteint la caractéristique «5V » du « Big Data» :

- **Volume** : l'énorme masse désordonnée d'informations circulant dans les medias sociaux

- **Vélocité** les activités des utilisateurs dans les RSNs sont imprévisibles, ce qui génère des données dynamiques et de manière très rapide, continue, devant être accessible à tout moment (en temps réel)
- **Variété** : les données sociales sont souvent bruitées et hétérogènes (combinaisons de texte, de liens, d'images, de vidéos et d'autres médias).
- **Véracité :** la véracité ou fiabilité des données est menacée par les comportements des internautes. Lorsqu'ils génèrent de fausses informations ou de faux profils.

- **Valeur :** les données sociales peuvent être très importantes, comme le cas de Cambridge Analytica qui aurait participé à la compagne électorale de Donald trump en diffusant notamment du contenu extrêmement ciblé aux utilisateurs Facebook [14].

1.5 Réseaux sociaux académiques

Ces dernières années, les réseaux sociaux académiques sont devenus inéluctables et s'imposent davantage au sein des communautés académiques (les scientifiques et les institutions de recherche). Ceci grâce aux fonctionnalités qu'ils offrent aux chercheurs en permettent de publier leurs résultats de recherche, de suivre l'activité scientifique, et de collaborer avec les différents individus du réseau. De ce fait simplifier et promouvoir les échanges entre scientifiques. Par ailleurs, ce type de réseaux est au profit des jeunes universitaires voulant intégrer le monde de la Recherche en leur permettant principalement de trouver des chercheurs experts, des articles pertinents, et étendre leurs recherches, Notamment d'être visible dans le monde de la recherche scientifique.

Tous les contenus sont librement accessibles, mais l'accès reste restreint (accès via login). Cependant, ces plateformes ne relèvent pas de l'Open Access [2] qui est basé sur l'absence d'obstacles (financier, juridique, techniquement) . [15] [16]

1.5.1 Exemple de plateformes académiques

Actuellement, il existe plusieurs plateformes exclusivement dédiées à la recherche scientifique, pouvant se présenter sous différentes formes du web dans le but de fournir leur datasets académiques pour aider les scientifiques à enrichir leurs recherches. Ces datasets sont des documents académiques

2. Un modèle de publication académique où les lecteurs ont accès aux œuvres publiées sans coût

intégrés qui contiennent de nombreux types de données numériques, beaucoup d'entre eux sont téléchargeables gratuitement [9]. Nous citons quelques exemples de ces plateformes :

- **Moteur de recherche** : Google scholar permet la recherche d'article en fournissant des métadonnées sur des documents scientifiques (articles, thèses et livres…etc.)
- **Base de données** : scopus est une base de données lacée par l'éditeur scientifique Elsevier. Elle correspond à un réservoir de document fournissant le contenu d'article.

- **Archive ouverte** : arXiv permettant les prépublications électroniques d'articles académiques dans les domaines de la physique, l'astrophysique, des mathématiques, de l'informatique, des sciences non-linéaires et de la biologie quantitative

- **Groupe d'édition** : springer représente un groupe éditorial et de presse spécialisée dans le secteur des Sciences, Technologies et Médecine.

- **Service d'information** : web of science signale la littérature scientifique mondiale en donnant accès à huit bases de données bibliographiques.

- **Site web** : ScienceDirect correspond à une plateforme gérée par l'éditeur Elsevier permettant d'accéder à plus des millions d'articles scientifiques.

- **Système d'identification** : ORCID permet de créer un identificateur unique pour chaque chercheur en le prémunissant contre les problèmes d'homonymie.

- **Réseaux sociaux académique** (RSAs) : tels Social Sciences Research Network, Nature Network, Mendeley. Néanmoins Ce sont ResearchGate et Academia.edu qui dominent le marché des RSAs. Nous présentons dans la table ci-dessous une brève description sur ResearchGate et Academia.edu, et différentes fonctionnalités offertes par ces réseaux, Ainsi que leur domaine de recherche et avantages

	ResearchGate	**Academia.edu**
Description	C'est un service de réseautage social qui a été fondé en 2008 par les physiciens Ijad Madish ,Sören Hofmayer, l'informaticien Horst Fickenscher. Le siège de l'entreprise ResearchGate est à Berlin [7].	Plateforme hébergée sur le Web pour les articles universitaires Lancé en 2008 par Richard Price, il est maintenu par l'entreprise Academia, dont le siège est à San Francisco en Clalifornie Le terme "edu" ne fait pas référence à une institution académique américaine, il s'agit par contre d'une entreprise privée [7].
Nombres d'utilisateurs	**+19 millions** de chercheur et scientifiques [8]	**114 millions** d'universitaires [13]
Nombres de publications	+**118 million** de publication [8]	**24 millions** d'articles [13]
Services proposé	-partager des publications et accéder à celle des autres -être visible dans le monde de la recherche -accéder à ses statistiques personnelles (savoir qui a lu, cité, recommandé son travail, ainsi que des détails tels le pays et les institutions d'où proviennent les lecteurs)[8] -collaborer avec ses homologues scientifiques ainsi qu'avec tout individu dans le réseau.[8] -adhérer à des groupes ciblés et accéder à des ressources liées au profit de la recherche.[8] -chercher du travail - un indice h et un score « RG score », ce sont deux façons de mesurer la réputation scientifique d'un chercheur et de refléter son activité sur ce réseau. En prenant compte ses interactions.[6] - possibilité de créer son propre réseau	-partager ses publications et accéder à celles des autres -être visible dans le monde de la recherche - surveiller des analyses approfondies autour de l'impact de sa recherche [13] -créer des pages spéciales ou les collègues et les pairs peuvent laisser des commentaires sur des articles ou des annotations spécifiques. [13] -tracer la recherche d'autres chercheurs [13]
Domaine	c'est les réseaux sociaux le plus utilisé en science exacte et science de la vie : biologie, Médecine, Informatique, Physique, Chimie [8]	physique, chimie, Biologie, Sciences médicales, Écologie, Science de la Terre, Sciences cognitives , mathématiques et informatique [13]
Forces	-sert fortement la documentation et la recherche scientifique en fonction du grand nombre de publication -utilise de grandes base de données, comme PubMed, arXIV, IEEE, RePEC et CiteSeer [8]	- favorise l'échange et la collaboration entre chercheurs vu le grand nombre d'utilisateurs dont il dispose .

Tableau 1-Academai VS ResearchGate [6] [7] [8] [13]

1.5.3 Topologie des RSAs

La structure d'un RSA devient de plus en plus complexe, vu la grandeur des données académique « Scholarly Big Data ». Cependant, le temps de calcul et la complexité augmentent en même temps. Pour cette raison, les chercheurs ont utilisé un graphe à base de sous-réseaux et ont développé un cadre généralisé de modèles de connectivité d'ordre supérieur. Le comportement social académique des chercheurs peut éventuellement changer au cours du temps. Ce qui qualifie les RSAs de réseaux dynamiques [3], et rend ainsi leurs structures topologiques difficiles à décrire. De ce fait, la modélisation des RSAs demeure très complexe. [9]

Les réseaux sociaux à caractère académique peuvent être construits dans diverses topologies selon les différences entres les nœuds du réseau et les relations entres eux. Nous constatons que les RSAs peuvent être divisés en deux grandes catégories. [9]

1.5.3.1 RSAs Homogènes

Les RSAs Homogènes font référence au réseau dont les nœuds représentent les mêmes entités, nous citons :

- **Réseau de co-auteurs**

C'est le fait que différents chercheurs collaborent entre eux afin d'effectuer une recherche puis rédigent les résultats sous la forme d'un document de recherche co-écrit. Les réseaux de co-auteur sont constitués principalement d'un ensemble d'auteurs, exemple dans la figure ci-contre auteur 1 et auteur 2 sont des co-auteurs de l'article 7 qui représente un document co-écrit (figure 2)

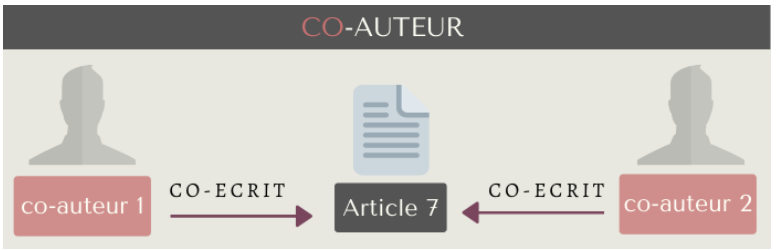

Figure 2-Co-Auteur

3. C'est un type de réseaux où les nœuds et les arrêtes peuvent apparaître ou disparaître.

- **Réseau de Co-citation**

La Co-citation désigne le fait qu'un même article cite deux publications différentes. Ainsi, les publications scientifiques représentent les nœuds fondamentaux dans ce type de réseau, exemple les articles 1 et 2 sont associés vu qu'ils sont co-cités dans la liste de référence Bibliographique de l'article 3 ce qui les (article 1 et article 2) définit comme Co-citation (figure 3)

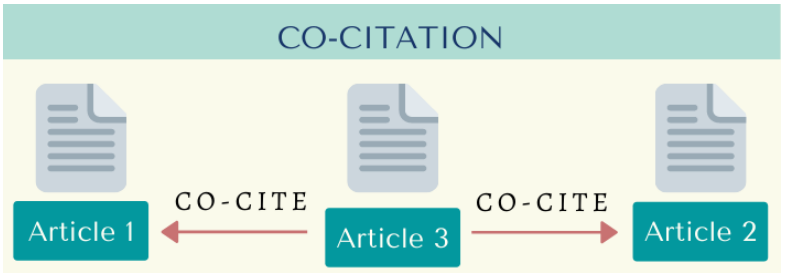

Figure 3-Co-Citation

- **Réseaux Co-word**

L'analyse de Co-mots représente la base des réseaux Co-mots, elle est largement utilisée à fin d'explorer le réseau de concept de sujet de recherche. Elle reflète la fréquence de co-occurrence de mot-clé qui désigne le nombre d'articles dans lesquels deux mots clés apparaissent en même temps.

1.5.3.2 RSAs Hétérogènes

Dans les RSAs Hétérogènes, les nœuds représentent des entités de type différent. Nous distinguons trois catégories :

- **Réseaux d'article-auteur**

Les réseaux d'article-auteur sont constitués de données hétérogènes (articles, auteurs, citations.etc.). Ils sont construits principalement dans le but d'analyser les liens entre les articles et les chercheurs, ce qui sert fortement la recommandation des articles appropriés aux chercheurs cibles.

- **Réseau de couplage bibliographique**

Un article scientifique peut être défini comme une unité de base du couplage entre deux publications Lorsque celles-ci le citent dans leur liste de référence bibliographique. Ce qui fait deux articles différents peuvent être lié en fonction du nombre de références communes entre eux, qui est

nommé la fréquence de couplage. Un exemple les articles 4 et 5 sont bibliographiquement couplés. (Figure 4)

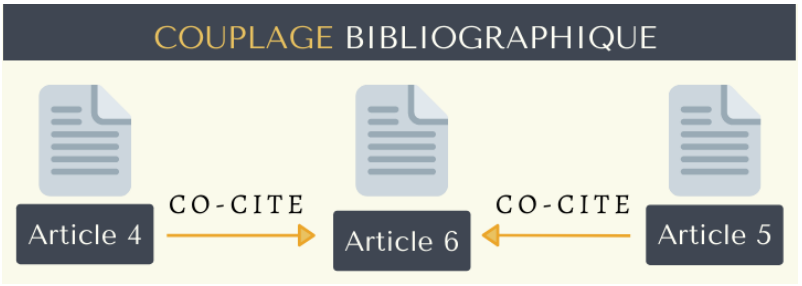

Figure 4-Couplage bibliographique

- **Réseaux Hybrides**

Cette catégorie de réseau est principalement utilisée dans l'identification des "research topics" ou des sujets de recherche. Les chercheurs ont proposé une nouvelle approche hybride pour l'analyse des revues, qui consiste à combiner la citation (sous la forme d'une "term-by- document matrix" qui désigne une matrice de terme par document) et le texte, (sous la forme d'une "cited-references-by-document matrix" qui désigne une matrice de références cité par document)

1.6 Conclusion

Dans ce premier chapitre, nous avons étudié de façon générale les réseaux sociaux numériques. En spécifiant l'intérêt de l'analyse de ces réseaux. Enfin par caractère cumulatif nous avons constaté que les RSNs, notamment les RSAs sont désormais l'usage principal du web. L'objet du second chapitre portera sur notre état de l'art : la modélisation des profils thématiques dans les RSAs.

chapitre 2

Etat de l’art

2.1 Introduction

Ce chapitre est consacré à notre état de l’art. En premier lieu, nous commençons par une brève définition de la représentation vectorielle des textes. En deuxième lieu, nous entamons le concept de la modélisation thématique, en définissant les différents aspects des modèles thématiques. En troisième lieu, nous passons au profil Utilisateur. D’abord, nous allons le définir en spécifiant son contenu. Après, nous décrivons les différentes phases de construction du profil.

2.1 Représentation vectorielle des textes

Afin d'effectuer des traitements sur le langage, une représentation mathématique des mots ou des documents serait nécessaire. Pour cela, de nombreuses méthodes ont été proposées depuis de longues années dans le but de représenter les documents. Cette tâche était traditionnellement réalisée grâce à la représentation du corpus dans le modèle d'espace vectoriel VSM de Salton, basée dans un premier temps sur la fréquence des mots. Le modèle consiste en fait à convertir un document en un vecteur composé de fréquences de mots, ainsi le corpus est représenté par une matrice document-terme ou chaque ligne correspond à un document du corpus et chaque colonne à un terme. VSM fut amélioré par la suite en lançant le système de terme_frequency inverse_docuement_frequency (TF-IDF), dans la mesure où chaque élément de la matrice document-terme est pondéré par la fonction (TF-IDF). L'idée principale de la pondération (TF-IDF) se base sur le fait que les termes qui apparaissent fréquemment dans un document particulier et répartis sur peu de documents prennent plus de poids. Etant donné que le modèle VSM n'est doté d'aucune technique tenant compte le sens des mots, il peut être généralement nuisible à la signification sémantique du texte. [21]

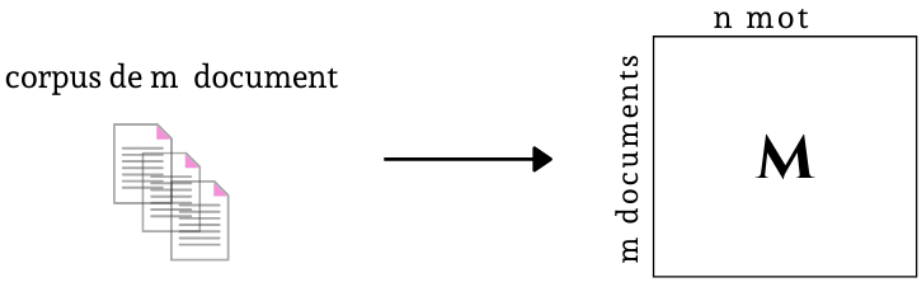

Figure 5-Représentation vectorielle de document

2.2 Modélisation thématique

La modélisation thématique peut être définie comme un processus d'apprentissage non supervisé, de reconnaissance et d'extraction de variables latentes (sujets) abordé dans un corpus de documents. Une approche proposée dans le but de réduire l'espace de la représentation vectorielle en résolvant les problèmes sémantiques de celle-ci [22] [21]. Plusieurs techniques furent développées dans ce sens y compris, l'analyse sémantique latente (LSA), l'analyse sémantique latente probabiliste (PLSA), et l'allocation de Dirichlet latent (LDA). Le principe détaillé de ces trois méthodes est donné dans les sous-sections suivantes.

2.2.1 LSA: L'analyse sémantique latente

L'analyse sémantique latente, ou en latent semantic analysis en englais (LSA), nommée aussi l'indexation sémantique latent et latent semattic indexation (LSI), une méthode mathématique et statistique proposée à la fin des années 80 dans le but de transformer une matrice document-terme en une matrice document-sujets. LSA étend le modèle d'espace vectoriel en utilisant une décomposition en valeurs singulières (SVD) à la matrice document-terme (**A**), suivie par une réduction de rang pour limiter le nombre de sujets latents extrait.

SVD transforme la matrice A en un produit de trois matrice A **= U * S * V^T** [23]:

- **U (m x t)**: matrice **document-sujets**, pour chaque document, une ligne (m) indique les sujets latents (t) qui sont présents dans le document (m).
- **V (n x t)**: matrice **terme-sujet**, pour chacun des t sujets latents la colonne associé de V indique les termes qui forment ce sujet.

- **S (t x t)**: les valeur sur la diagonale de S_t indiquent l'importance de chacun des t sujets latents.

2.2.2 PLSA: L'analyse sémantique latente probabiliste

2.2.2.1 Définition

L'analyse sémantique latente probabiliste PLSA est un modèle de sujet génératif probabiliste, dérivée de la perspective statistique de la LSA. La méthode PLSA fut proposée la première fois par thomas hoffman. Afin de répondre aux lacunes souvent critiquées de la LSA, à savoir qu'il y a une limitation concernant le manque d'interprétation probabiliste ou statistique. En effet, la PLSA améliore considérablement la précision par rapport à la LSA [24] [25]. Cependant, le paradigme PLSA a été largement utilisé dans de nombreux domaines de l'intelligence artificielle connexes à l'extraction d'information, filtrage de l'information, le traitement automatique du langage (TAL) et l'apprentissage automatique.

2.2.2.2 Principe du PLSA

L'approche probabiliste PLSA est basée sur le modèle d'aspect qui introduit l'hypothèse d'indépendance conditionnelle. Cependant, un document $d \in D = \{d_1,d_2,\dots d_M\}$ et un terme $w \in W = \{w_1,w_2,\dots w_N\}$ sont conditionnellement indépendants sachant un thème latent $z \in Z$. En effet, PLSA est un modèle à variable latentes, dans la mesure ou chaque variable cachée z est associée à des données co-occurrentes comme la distribution des termes w au sein d'un document d [24]. Le modèle est défini à l'aide de trois paramètres probabilistes: **P(d)** représente la probabilité de sélectionner un document d au sein du corpus, **P(w|z)** est la distribution du terme w dans le thème latent z, **P(z|d)** correspond à la répartition du thème (z) dans le document d. Cependant, ce modèle génératif peut être défini en trois étapes [24] [25]:

1. Sélectionner un document d basé sur la probabilité P(d).
2. Sachant le document d, le sujet z du document est sélectionnée selon la probabilité conditionnelle P (z|d).
3. Après avoir sélectionné le thème z, les termes w du document sont choisis selon la probabilité conditionnelle P (w|z).

Le duo de variables observées (d, w) peut être généré au moyen d'une conversion du processus génératif en une distribution de probabilité conjointe, la formule est la suivante [24] [25]:

$$\mathbf{P(d,w) = P(d)\ \sum_z P(w|z)P(z|d)}$$

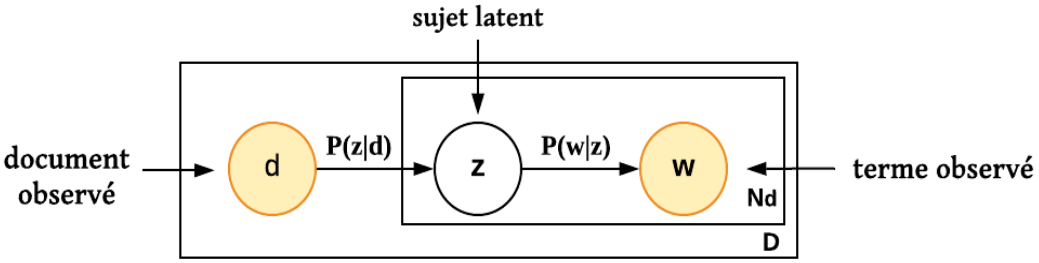

Figure 6-Graphe de fonctionnement de PLSA

2.2.3 LDA: allocation de Dirichlet latent

2.2.3.1 Définition

Le modèle LDA (allocation de Dirichlet latente) ou latent Dirichlet allocation en englais est une autre approche pour interpréter de grandes quantités de données textuelles et mesurer leur signification linguistique latente. Proposé par blei et al en 2003 pour résoudre les problèmes du modèle PLSA. En effet, LDA est un modèle hiérarchique avec une structure plus stable permettant d'éviter toute condition de surapprentissage, vu que son espace de paramètres est limité et n'augmentant pas avec la taille du corpus. Par ailleurs, le paradigme LDA est considéré comme un modèle probabiliste génératif complet, dans le sens où il permet une bonne estimation de la probabilité d'un document non rencontré lors de la phase d'entraînement [26]. LDA est l'une des techniques de modélisation de sujets les plus populaires qui trouve des applications dans de nombreux domains: l'analyse de texte et la TAL, la vision par ordinateur, système de recommandation l'analyse de réseaux sociaux. [27]

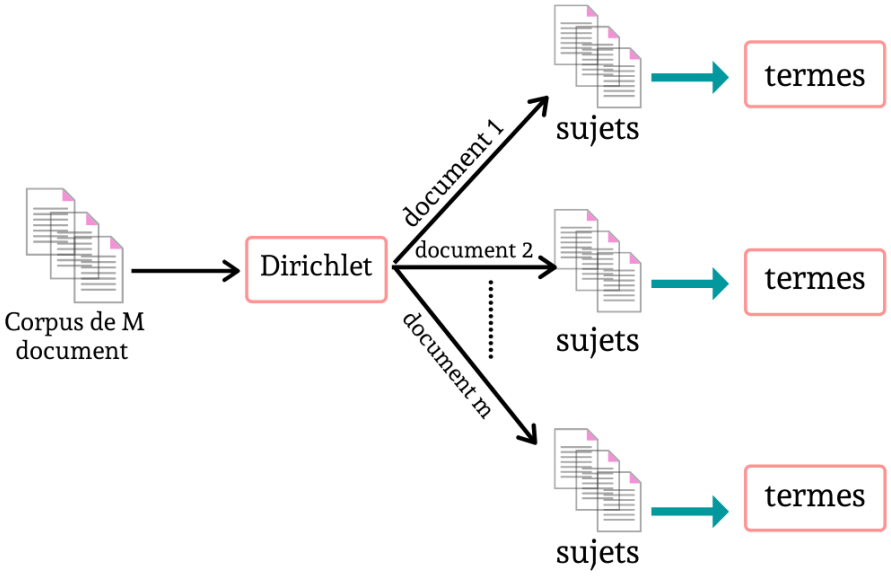

Figure 7-LDA Entrées / sorties

2.2.3.2 Principe du LDA

Le modèle LDA a adopté une approche différente de la PLSA en s'appuyant sur une loi de Dirichlet a priori pour les sujets latents. L'idée principale de cette méthode se base sur le fait que chaque document est modélisé comme un mélange de sujets et chaque sujet est une distribution sur des mots. Après avoir appliqué LDA sur un corpus de document, deux résultats principaux se présentent, une distribution de document-sujet et une distribution de termes-sujets [26][28]. La figure ci-contre met en évidence les entrées-sorties du modèle LDA.

Avant de définir formellement le processus LDA, nous introduisons quelques notations [28]:

- D représente le corpus de document.
- K fait référence au nombre de thèmes composant le modèle. K est supposé connu au préalable et fixe.
- Chaque sujet Φ_k, ou $1 \leq k \leq K$ est une distribution sur un vocabulaire fixe de termes et $\Phi_{k,w}$ est la proportion du terme w dans le sujet k.
- θ_d est le mélange de sujets dans le document d, et $\theta_{d,k}$ est la proportion du sujet k dans le document d.
- z_d indique les affectations de sujet pour le document d, où $z_{d,n}$ est le sujet assigné au nième terme dans le document d.
- w_d est le terme apparaissant dans le document d, où $w_{d,n}$ est le nième terme dans le document. Tous les termes sont des éléments d'un vocabulaire fixe.
- β est le Dirichlet a priori des distributions sujet-terme.
- α est le Dirichlet a priori sur les distributions de document-topics.

Le processus de cette technique generative peut être défini comme suite [26] [28] :

1. Pour chaque sujet k, tirer une distribution sur les mots des thèmes Φ_k à partir d'une distribution Dirichlet avec paramètre β, ainsi multinomiale $\Phi_k \sim Dir(\beta)$ ou $1 \leq k \leq K$
2. Pour chaque document d,
    a. Tirer une distribution de proportion de thème sur le document d multinomiale $\theta_d \sim Dir(\alpha)$.
    b. Pour chaque mot $w$du document d,
        i. Choisir un sujet $z_{d,n}$ au hasard à partir de la Distribution θ_d pour le nième terme dans le document d, où $z_{d,n} \sim Multinomial(\theta_d)$.
        ii. Ensuite, Sachant le theme $z_{d,n}$ préalablement choisi et le mot w, sélectionner le mot le plus probable $w_{d,n} \sim Multinomial(\Phi z_{d,n})$.

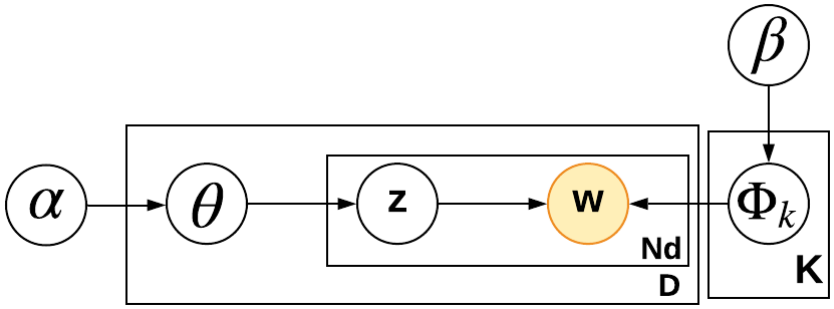

Figure 8-Graphe de fonctionnement de LDA

Depuis les années 80, les modèles thématiques ne cessent d'évoluer, la figure ci-dessous présente le développement chronologique de ces modèleés ainsi que leurs divers avantages et inconvenient

.

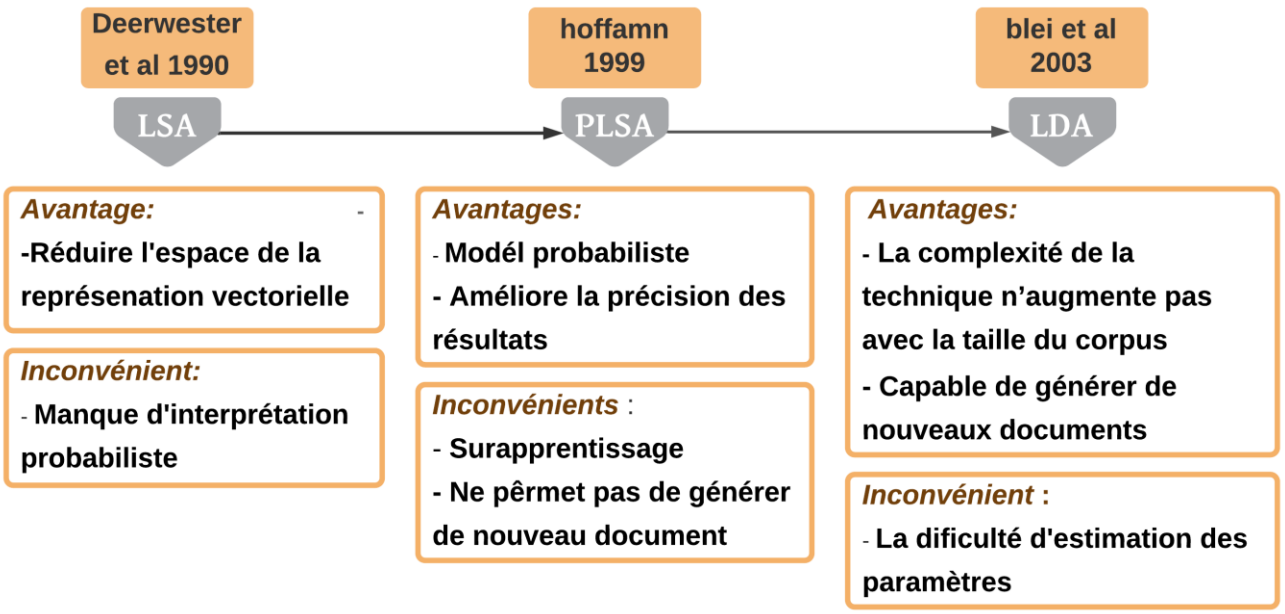

Figure 9-Avantages et inconvénients des modèles thématiques

2.3 Profil utilisateur

Un profil d'utilisateur peut généralement être décrit comme une collection d'informations personnelles qui caractérisent un utilisateur et reflètent ses préférences, ses objectifs, ses besoins, ses comportements et son intérêt. Les termes profil d'utilisateur et modèle d'utilisateur paraissent exactement identique, mais il existe une différence entre eux ou les profils servent exclusivement de source d'informations qui n'incluent pas les interactions entre les utilisateurs et un système. Les modèles en contre partie contiennent de telles informations, ce qui permet à l'environnement des RSNs de s'adapter à l'utilisateur, ce qui fut indispensable pour tout système de personnalisation. D'autre part le type d'information qui forme un modèle d'utilisateur varie considérablement, selon le domaine d'application utilisé tels les informations qui circulent sur les plateformes de RSNs Généralistes par rapport aux RSAs [29][30]. Dans les RSAs, les membres se concentrent principalement sur des informations dans le contexte de la recherche scientifique et pour cela le modèle d'un chercheur inclut les mêmes informations qu'un modèle d'utilisateur en générale (données personnelles, Interaction, intérêts) avec quelques particularités. Le contenue détaillée de chaque modèle est donnée dans le tableau ci-dessous.

Un utilisateur en général		Un chercheur	
Données personnelles	Nom , Prénom , Age , Adresse…	Communication	(Conférences invitées, séminaires, etc)
Interaction	L'ensemble des informations qui décrit le comportement de l'utilisateur (son historique , les annotations , ses préférences) [32]	Formations académiques ou non académiques	(Nom du titre obtenu (Master, PhD, etc), date d'obtention, nom de l'institution, ville, pays)
intérêts	Exprime son domaine d'expertise ou son périmètre d'exploration [32]	intérêts	Dans le contexte de la recherche scientifique un intérêt correspond à un article, livre ou une publication à laquelle s'intéresse le chercheur ou son domaine de recherche

Tableau 2-Contenue d'un modèle utilisateur en général et un modèle d'un chercheur

2.3.1 Profilage

Après avoir donné un bref aperçu des profils et des modèles utilisateur dans la section précédente, l'accent sera désormais mis sur les méthodes de profilage, qui peut être défini comme le processus de création du modèle d'utilisateur, Il sert a évaluer et identifier les intérêts, traits et caractéristiques personnels d'un utilisateur. Plusieurs approche d'intelligence artificielle peuvent être utilisées dans ce sens, tels que les réseaux bayésiens, les algorithmes génétiques, ou les réseaux neuronaux. En outre, le profilage est l'un des éléments de base dans de nombreux domaines, tels que l'analyse des RSNs, les systèmes de recommandation, bien qu'il ait principalement évolué par le biais du datamining et de l'apprentissage automatique [30]. La technique de profilage se compose généralement de trois phases principales la première concerne l'acquisition de données pertinentes sur les utilisateurs individuels. La deuxième phase se concentre sur la construction du modèle utilisateur à partir des données collectées. La troisième cible la mise à jour du modèle utilisateur [31].

La description détaillée de chaque processus est donnée dans les sous-sections ci-dessus.

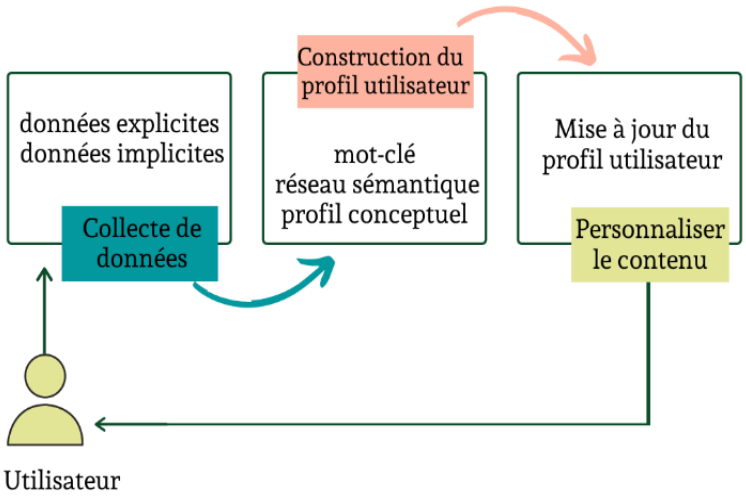

Figure 10-Les phases du profilage

2.3.2 Collecte de Données

La collecte d'informations sur un utilisateur particulier est le point de départ des techniques de profilage. Cette phase consiste principalement à acquérir les informations sur l'internaute. Cependant, il est prévu que le système identifie exclusivement les utilisateurs, ce qui constitue l'exigence essentielle du système. Pour cette tâche, il est possible d'obtenir les informations sous la forme d'entrées de l'utilisateur ou de les collecter automatiquement par un agent intelligent. Ainsi, deux approches de collecte d'information se présentent explicite et implicite [31] [29] [32] [30] .

2.3.2.1 Méthode explicite

La collecte explicite d'informations est une technique simple, qui repose sur les informations fournies directement par l'utilisateur. Dans ce sens, le modèle utilisateur peut être obtenu en demandant à l'utilisateur de remplir un formulaire pouvant contenir son nom, son adresse, son numéro de téléphone, son statut professionnel, son anniversaire, etc. Ce type de formulaire peut également permettre à l'utilisateur d'exprimer son opinion ou spécifier ses intérêts et ses préférences. Cependant, les informations explicites sont fournies par les utilisateurs via un processus d'enquête et d'enregistrement. [30] [31]

2.3.2.2 Méthode Implicite

La technique de conception implicite contraste fortement avec l'approche explicite. Vu que le système doit déduire par lui-même les informations permettant de construire le modèle utilisateur. La méthode implicite se base principalement sur les interactions, le comportement et les activités des utilisateurs au sein du système, elle implique également le contenu des modèles des individus similaires à l'utilisateur et les individus dans le réseau social de l'utilisateur. En raison de la complexité accrue de cette approche, il est nécessaire de déployer des techniques fortement développées appartenant à divers domaines de recherche et académiques tels que l'intelligence artificielle, l'apprentissage automatique et la science sociale. [30] [29] [31] [32]

Les deux approches implicite et explicite ont leurs forces et leurs faiblesses comme mentionnées ci-dessous:

	Méthode Explicite	Méthode Implicite
Avantage	-simplicité de mise en œuvre [30]	- mises à jour automatiques (les nouvelles entrées sont automatiquement traitées et les informations résultantes attribuées) [30] -aucune dépendance à la volonté des utilisateurs pendant le processus de construction [30][31]
Inconvénient	- dépend fortement de la volonté des utilisateurs [29] [30] - nécessite la fiabilité de l'utilisateur (les informations peuvent être frauduleuses ou erronées) [32] - profil à caractère statique (les utilisateurs doivent s'assurer par eux même que les descriptions sont mises à jour à tout moment) [32] [30]	- nécessite une grande quantité d'interaction dans la phase initiale avant la création d'un modèle utilisateur précis [31] - complexité de mise en œuvre [30]

Tableau 3-Avantages et inconvénients des méthodes de collecte de données

2.3.2.3 Prétraitement de données

Les données textuelles potentiellement utiles dans un projet de data mining peuvent être générées quotidiennement en temps réel par les innombrables activités auxquelles nous nous connectons et participons en ligne chaque jour. Cependant, la majorité de ces données se présentent comme une forme hautement non structurée. Les informations contenues dans ces données ne sont pas exploitables directement, ils peuvent se présenter sous forme de: informations redondantes, données incomplètes ou bruyant (contenant des valeurs aberrantes). C'est ici que se porte tout l'intérêt du preprocessing, l'une des étapes fondamentales du processus d'exploration de données, son rôle principale consiste à convertir les données brutes en un ensemble de données propres, exploitables et analysables. [18] [19] [20]

2.3.2.3.1 Techniques de prétraitement de données

Il existe différentes formes de prétraitements de données, les sous-sections suivantes présentent les plus efficaces.

- **La tokenisation**

l'analyse d'un texte démarre généralement par une étape de segmentation de mots ou tokenisation. Un processus de décomposition du texte, en unités plus petites appellées jeton ou token *(*ex, un mot, un chiffre, un signe de ponctuation, ou plusieurs mots). Cependant, la technique de tokenisation est souvent effectuer en tenant compte des régles dépendant de la langue et du lexique. [17] [18] [19]

- **Suprression du bruit**

La suppression du bruit ou « Noise Removal » en anglais est l'une des premières étapes du prétraitement de données, qui correspond à la suppression d'entités bruyantes, pouvant nuire a l'analyse du texte et produire des résultats incohérents (ex, ponctuation, caractères spéciaux, etc). C'est l'une des étapes les plus essentielles concernant l'exploration de texte et la TAL. [19]

- **Supression des Stopwords**

Le filtrage des mots vides ou des mots d'arrêt fait référence à la suppression des mots fréquents (ex, à, et, comme, le, un, etc) généralement présents dans tous les textes, qui pour certains modèle d'apprentissage fournissent peu d'informations, ou dans certains cas introduisent un bruit inutile et doivent donc être supprimés. [17] [19]

- **Normalisation**

Après avoir scinder le texte en token et nettoyer les données, la question suivante est de savoir quels token sont vraiment distincts. Il est purement nécessaire de traiter comme un token unique les différentes variantes issus d'une meme forme canonique, afin de se baser sur le fond thématique. Cela peut être réalisable grâce aux techniques de normalisation suivantes:

  - **Lowercasing**

Cette étape consiste à la mise en minuscule de toutes les données de texte, c'est l'une des formes de prétraitement de texte les plus simples et les plus efficaces. [17]

  - **La lemmatisation**

C'est un processus d'extracion de forme de base, radical ou « lemme » des mots, visant à supprimer les terminaisons flectionnelles en utilisant le vocabulaire et l'analyse morphologique. Ce processus permet en fait de convertir les mots à leur racine réelle. [17] [18] [19]

  - **La Stemmatisation**

La Stemmatisation, tige, racinisation ou « stemming » en englais, se base sur le meme principe d'extraction de forme de base, que celui de la lemmatisation. Néaumoins la racine ou le « stemme » extrait n'est pas nécessairement un mot de la langue. [17] [18] [19]

Donnée brut	Noise Removal	Lowercasing	Stop-words	Lemmatisation	Stemming
<a> A CheRché < /a>	A CheRché	a cherché	cherché	chercher	cherch
1.Ne pAs chErcher	Ne pAs chErcher	ne pas chercher	chercher		
& !Un CHEVAl..	Un CHEVAl	un cheval	cheval	cheval	cheva
//?Des CHevauX%	Des CHevauX	des chevaux	chevaux		

Tableau 4-Exemples de quelques techniques de prétraitement

- **Standarisation**

Une étape clé de prétraitement, désigne un processus de transformation d'un texte en une forme canonique (standard). Cela concerne les phénomènes orthographiques qui ne sont pas reconnus par les modèles d'apprentissage et les moteurs de recherche. Dans les médias sociaux ce type de données apparaît fortement tel l'allongement expressif (cooooool) et les mots qui ne font pas partis des dictionnaires lexicaux standard (ex, prsk, prk, mrc). [17] [19]

- **Part of speech (POS)**

Dans de nombreux types de textes, le sens réel des mots peut être perdu après la tokenisation. Cela est dû à la variation de sens que les mots peuvent régulièrement avoir en fonction de leur utilisation. La partie du balisage vocal ou la POS définit la fonction et l'utilisation des mots a l'intérieur des phrases, dans la mesure ou chaque mot est associé à une POS (adverbes, verbes, noms, adjectifs). Le balisage POS est largement utilisé afin d'améliorer diverses techniques de prétraitement (tokenisation, normalisation, stopwords) et les appliquer d'une manière plus approfondi. [19]

Téchnique	Rôle	Rôle commun principal
tokenisation	Permet le traitement de petites entités (token) au lieu de traiter la totalité du texte doté d'un espace de grande dimenssion .[17]	Limiter la taille du vocabulaire, ce qui réduit la mémoire des modèles d'apprentissage et augmente la vitesse de prédiction. [17]
Nettoyage de données: ▪ Noise removal ▪ stopword	Supprimer les entités qui ne sont pas pertinentes pour le contenue des données.[17][19]	
Normalisation: ▪ Lowercasing ▪ Lemmatisation ▪ stemming Standarisation	Réduire la variabilité des entités issus d'une même forme canonique.[17][18][19]	
Part of speech (POS)	Permet une désambiguïsation efficace du sens des mots. conservez le contexte des mots. [19]	

Tableau 5-Rôle des techniques de prétraitement des donnée

2.3.3 Construction du modèle utilisateur

Dans la construction du modèle utilisateur, divers algorithmes d'apprentissage et systèmes de récupération d'informations sont utilisés en fonction du choix de la représentation. La construction du modèle peut être sous forme d'un ensemble de mots-clés pondérés, de réseau sémantique ou d'un profil conceptuel. Le principe de chaque approche est donné ci-dessous [31] [32] [33]:

2.3.3.1 Mots clés

Les modèle basés sur des mots-clés font partie des premières représentations du modèle utilisateur. Ils sont construits en extrayant les mots-clés des pages web dans les historiques de recherche. Chaque mot extrait se voit également pondéré selon la technique de pondération choisie, la plus souvent utilisé se base sur la fonction TF-IDF proposé par Gérard Salton [21]. Cependant, le but de cette pondération est de préciser l'importance de chaque intérêt par rapport aux autres intérêts de l'ensemble. D'autre part un mot clé peut simplement correspondre à un intérêt (ex. rock, rap, football) ou à un domaine

d'intérêt (ex, musique, sport). Ainsi, les intérêts de l'utilisateur peuvent être sous forme d'un seul vecteur de termes extraits où chaque mot clé correspond à un intérêt, ce qui ne représente pas le véritable comportement de l'utilisateur. En outre dans une approche plus précise les intérêts peuvent être structurés en plusieurs vecteurs ou chaque vecteur représente une catégorie d'intérêt.

2.3.3.2 Réseau sémantique

La représentation basée sur réseau sémantique, est similaire à la représentation par mot clé, du fait que ces deux méthodes se basent sur la même idée d'extraction de terme pondéré. Néanmoins, la structure des terme extrait représente toute la différence entre ces deux techniques de construction. Vu que les modèles utilisateurs issus de la représentation par réseau sémantique sont créés en ajoutant les mots clés à un réseau de nœuds plutôt qu'un vecteur. Dans la mesure ou chaque nœud représente un concept, Ce type de modèle est plus précis que celui basé sur des mots clés vu qu'il élimine les problèmes d'ambiguïté, de synonymie et de polysémie.

2.3.3.3 Profil conceptuel

L'approche de construction du modèle utilisateur basés sur les concepts, repose sur le même principe de nœuds, que celui de la représentation par réseau sémantique. Cependant, dans la représentation conceptuelle, les nœuds représentent un sujet abstrait au lieu des termes spécifiques ou des ensembles de mots associés, telle la représentation par réseau sémantique. Les modèles conceptuels sont également similaires aux modèle basé sur mots-clés, dans le sens où chaque concept décrivant un intérêt est structuré sous forme d'un vecteur de caractéristiques pondérées, ainsi une valeur numérique ou un poids est attribué afin d'exprimer le degré d'intérêt de l'utilisateur pour chaque concept. Ce type de modèle est représenté sous forme d'une hiérarchie de concepts pondérés, au moyen d'une association des intérêts de l'utilisateur aux concepts existants en taxonomie. Si le même concept apparaît de nouveau, alors son poids est augmenté de un.

2.3.4 Mise à jour du modèle utilisateur

La mise à jour du modèle utilisateur est un processus, qui se déroule souvent suite à la création réussie du modèle utilisateur. Cette tâche est tout à fait nécessaire afin de permettre au system de s'adapter à l'utilisateur, notamment à la variation de ces intérêts. Comme le cas des RSNs ou l'évolution des intérêts présente un défi majeur lors de la personnalisation du contenu des utilisateur, vu que les préférences des individus peuvent changer au cours du temps. D'autre part la mise à jour du

modèle intervient après chaque soumission de requête ou le système identifie la cible et son occurrence dans modèle. Afin de fournir à l’utilisateur des services utiles et personnalisé. [31] [33]

2.3.5 Structure d’un article scientifique

Chaque chercheur, doit déclarer les résultats de ses recherches. Ces informations doivent réspecter certaines contraintes (la clarté, la validité, la fiabilité...etc). Dans le but de répondre a ces exigences, ce dernier doit être préparé de manière appropriée suivant une structure spécifique qui comprend : [34]

- **Le titre**

L'idée de base de la recherche est décrite principalement en une phrase. Cependant, le titre de l'article crée la première impression sur l'article lui-même, en particulier, sur la recherche effectuée.

- **Le résumé**

Un résumé est constitué de courtes caractéristiques d'un article scientifique, qui permettent de décrire la recherche effectuée, tel le but, le contenu, le type, la forme et d’autres particularités de l’article qui sont prises en considération également. Un résumé bien rédigé offre des possibilités que la recherche ne soit pas mal utilisé ou ignoré. Vu qu'il s'agit d'un texte de taille très limitée, dans lequel toute l'essence de l'article doit être reflétée.

- **Les mots clés**

À la fin du résumé, des mots-clés sont généralement présentés. Le plus souvent, de 3 à 5 mots-clés sont présentés. À l'exception de l'idée principale de recherche (problème), au moins un mot-clé doit décrire la méthode de recherche effectuée. L'auteur d'un article doit choisir les mots-clés appropriés qui caractérisent au mieux la recherche décrite.

Figure 11-Structure d'un article scientifique

2.4 conclusion

Dans ce chapitre nous avons abordé essentiellement les notions de base de la modélisation thématique, ainsi que ceux du profil utilisateur. l'objectif des modèles de thème dans notre travail est d'extraire les thèmes qui intéressent chaque chercheur (ses intérêts), Ce qui nous permettra de modéliser les profils thématiques des chercheurs, Cette tache fera l'objet du prochain chapitre .

chapitre 3

Approche proposée

Partie: I

3. I.1. Introduction

Nous allons présenter dans la première partie du troisième chapitre l'approche que nous avons proposées pour modéliser thématiquement des profils de chercheur dans les RSAs. Dans la mesure où nous allons représenter chaque chercheur par son domaine d'intérêt à l'aide des articles scientifique qui l'intéressent, les diverses phases de notre approche sont citées en détaille dans ce qui suit.

3. I.2 Description générale de l'approche proposée

Notre travail est structuré en trois étapes essentielles (Figure 12) pour modéliser les profils des chercheurs, la première concerne l'acquisition des données, ensuite nous effectuons différentes techniques de bag of word (pré-traitement) sur le corpus. Après, nous avons établi la vectorisation sur nos données prétraitées. Une fois nos données vectorisées viens la phase de formation de notre modèle d'apprentissage. Cependant, nous appliquons LDA sur nos données vectorisées à fin d'extraire les topics. Enfin, arrivent l'évaluation et la validation de notre modèle, celle-ci nous permettra de choisir le modèle le mieux interprétable

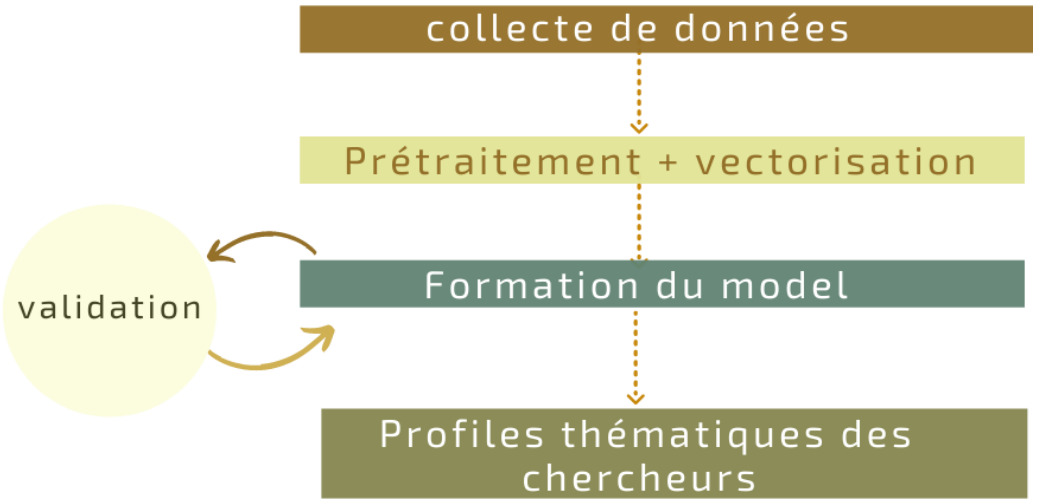

Figure 12-Etapes de l'approche proposée

3. I.2 .1 Etape 1: collecte de données

4.

Avant que notre étude puisse commencer, les données appropriées doivent être extraites. De ce fait les différents articles reflétant les intérêts de chaque chercheur sont capturés à l'aide de la technique du web crawling, ces articles en effet peuvent être issus des marquages du chercheur, ses partages, ses téléchargements, et ses recommandations au sein de la plateforme de son réseau social académique. Ainsi, un corpus de document est créé pour chaque chercheur. Enfin, l'objectif principal de cette phase arrive, qui est la construction d'un jeu de données (dataset). Le dataset avec le quell nous avons travaillé a été téléchargé à partir de Mendeley Data [4] . Nous avons créé deux datasets sous format excel à partir de dataset principal qui contient environ 35387 articles scientifiques, contenant chacun l'identifiant du chercheur, le nom, le titre, les mots-clés et le résumé de chaque article. Le dataset n°1 contient 1000 chercheurs ayant un seul document chacun, tandis que le deuxième contient 2000 chercheurs ayant plusieur documents dans leur corpus (Figure 13)

[4] https://data.mendeley.com/datasets/zm33cdndxs/draft?a=ef6d0d03-1102-4b7e-9195-43dd9d1be3b1

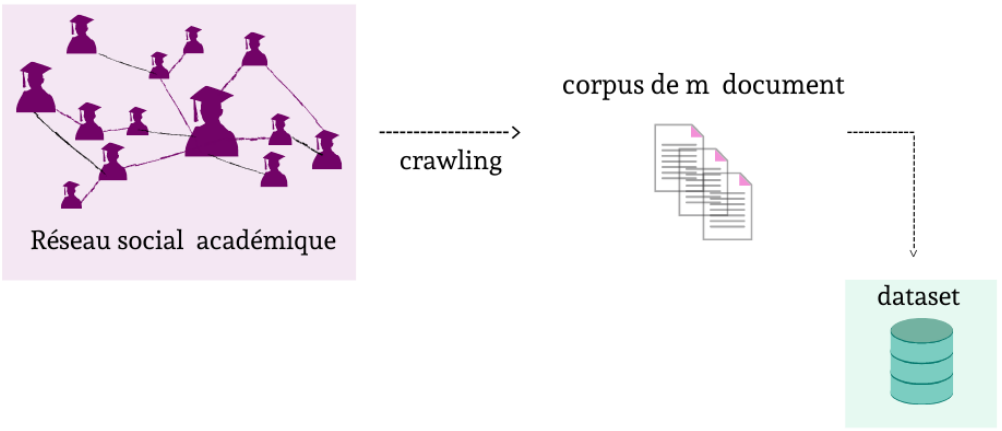

Figure 13-

Collecte de données (crawling)

3. I.2.2 Etape 2: prétraitement

Cette étape est cruciale dans toute approche d'exploration de corpus textuel, elle consiste à rendre les données collectées (dataset), exploitable en supprimant le bruit tel que les ponctuations, les caractères spéciaux et les mots qui ne sont pas porteur de sens. Après cette phase, chaque document sera représenté par un vecteur de la taille du vocabulaire, on parle de représentation en sac de mots (bag of word). en pratique, cette étape produit la matrice term_docuement _frequency (figure 14)

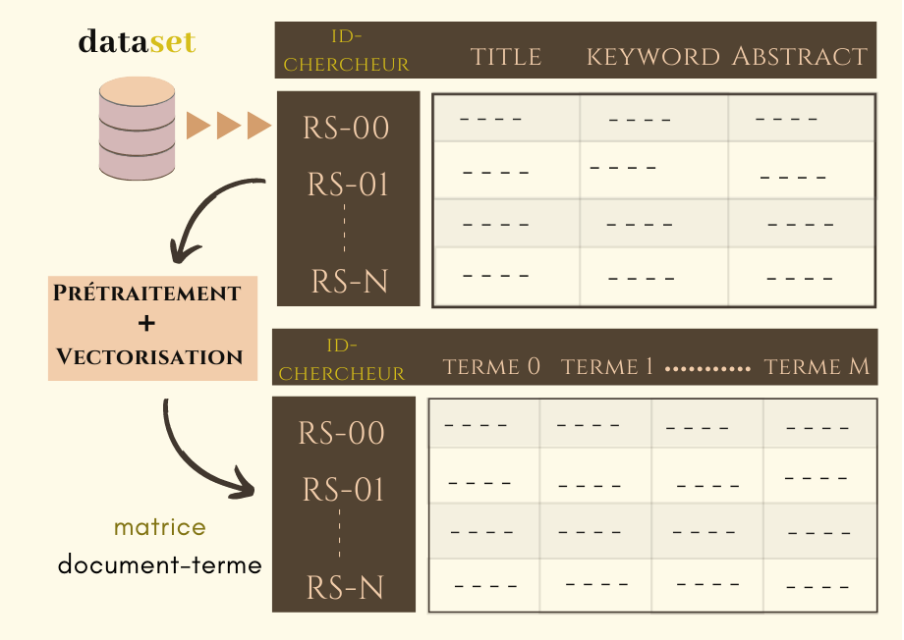

Figure 14-Prétraitement et Vectorisation

3. I.2.3 Etape 3: l'entraînement du modèle LDA

Cette phase concerne principalement la formation de notre modèle d'apprentissage. Elle consiste a crée une matrice document-topic, contenant les proportions des topics pour chaque document. Ensuite, nous avons crée une matrice chercheur-topic contenant les distributions des topics pour chaque chercheur. Comme le montre la (figure 15).

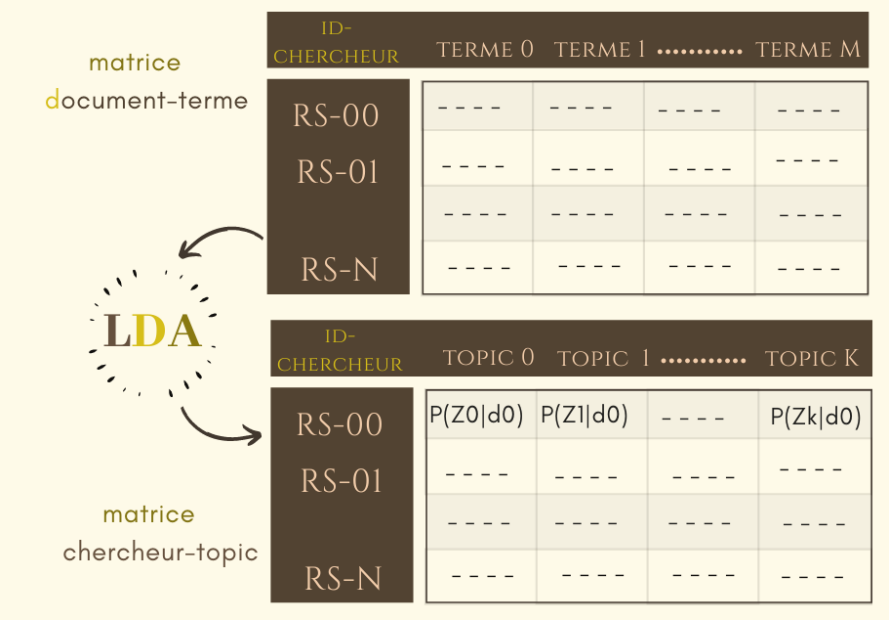

Figure 15-Création de la matrice chercheur-topic

L'évaluation de la qualité (interpérabilité des topics) des modèlees LDA, reste un problème critique à ce jour. Les chercheurs du domaine de TAL et de la recherche d'information se sont divisés en deux groupes. Le premier préconise de découper les documents du corpus en deux parties (données entraînement et données de test) [35] [36]. Tandis que le deuxième utilise l'ensemble du corpus pour évaluer la pertinence du modèle [37] [38], en s'appuyant sur des métriques d'évaluation comme la cohérence. Les scores de cohérence évaluent la qualité des sujets en calculant le degré de similitude

sémantique entre les mots principaux de chaque sujet. Un modèle LDA dont la cohérence est élevée infère des topics de bonne qualité (les mieux intérprétables) [39].

3. I.3 Scénario de modélisation

Notre étude consiste à modéliser les profils thématiques des chercheurs dans un réseau social académique. Pour la réaliser, nous avons établi une étude quantitative en distinguant deux cas principaux (Figure 16), le premier vise à modéliser les chercheurs juniors ayant un seul document. Tandis que le deuxième se concentre sur la modélisation des chercheurs seniors ayant un corpus Q de m documents tel que Q = {d1, d2,……, dm}.

Figure 16-Etude des cas

3. I.3.1 Cas 1: Chercheur junior

L'idée de base de la modélisation des profils thématiques des chercheurs a l'aide de LDA, est de représenter chaque chercheur par un vecteur de probabilités $\vec{f}$, dans la mesure où chaque composant est associée à un thème k (allant de 1 à K). Et contient la valeur de probabilité du sujet k étant donné le chercheur (Figure 17).

$$\vec{f} = \langle P(topic_1), P(topic_2), P(topic_3) \ldots\ldots\ldots, P(topic_k) \rangle$$

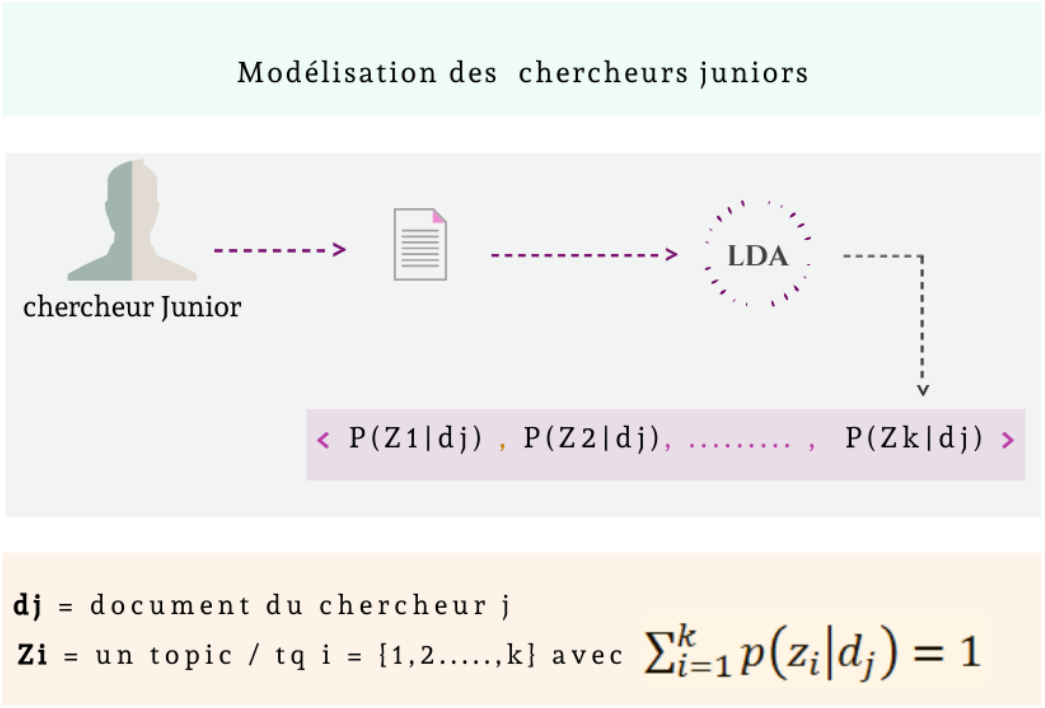

Figure 17-Modélisation des chercheurs juniors

Dans cette partie nous avons varié le nombre de sujets, dans le but de trouver le modèle le mieux interprétable qui nous permettra de modéliser les profils des chercheurs. En plus, de la première étude (quantitative), nous avons établi une étude qualificative en formant quatre modèle LDA (figure 18), a partir des trois caractéristique (title, keyword, abstract) de notre jeu de données. Le principe et les résultats obtenus dans chaque modèle sont cités en détail dans ce qui suit.

1. **Modèle 1** : ⟨ Title , Keywords , Abstract ⟩
2. **Modèle 2** : ⟨ Title , Keywords ⟩
3. **Modèle 3** : ⟨ Title , Abstract ⟩
4. **Modèle 4** : ⟨ Keywords , Abstract ⟩

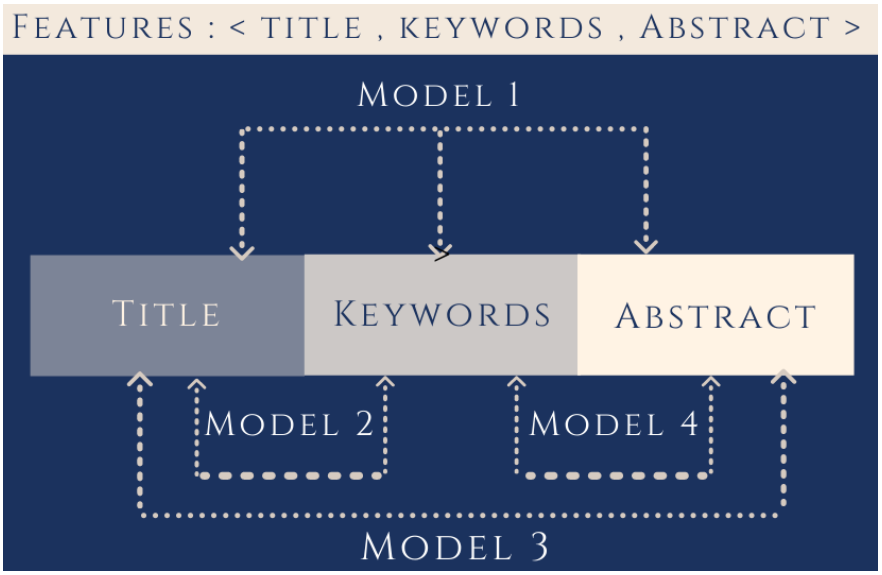

Figure 18-Les quatre modèles LDA

Le tableau ci-dessous montre les résultats (cohérence de sujets) issus des quatre modèles

	Nombre de topic	Modèle n°1	**Modèle n°2**	Modèle n°3	Modèle n°4
Cohérence	5	0.3489	0.4932	0.3685	0.3632
	10	0.376	0.522	0.375	0.3623
	15	0.3633	0.5476	0.378	0.3686
	20	0.371	0.5497	0.3728	0.3803

Tableau 6-Résultat obtenue dans le premier cas

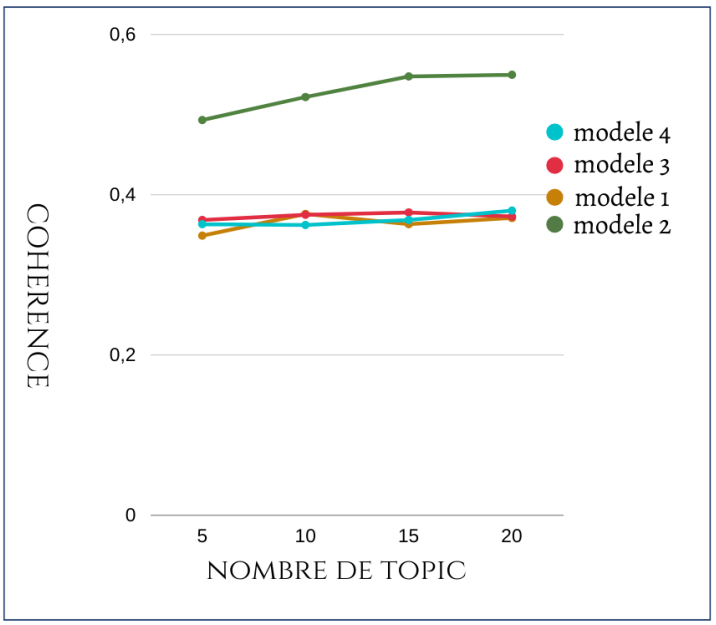

Figure 19-Etude comparative / Cas 1

3. I.3.2 Cas 2: Chercheur senior

Ce deuxième cas repose sur les mêmes principes de création de quatre modèles d'apprentissage et de variation du nombre de topic que ceux du premier cas. Néanmoins, la manière de modéliser les chercheurs est tout à fait différente. Vu que nous avons appliquer LDA au corpus des chercheurs ayant plusieurs documents. Cependant, le profil d'un chercheur est construit en agrégeant les distributions des sujets dans tous les articles contenant dans son corpus. Au moyens de deux fonction la moyenne (eq.**1**) et le maximum (eq.**2**) comme le montre la figure ci contre. (figure 20)

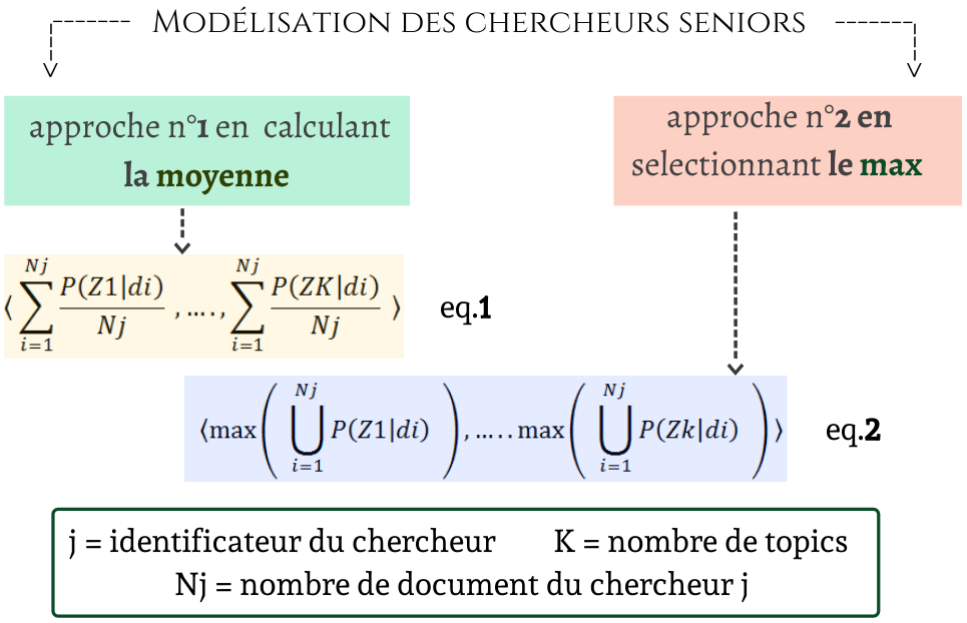

Figure 20-Modélisation des chercheurs seniors

	Nombre de topic	Modèle n°1	**Modèle n°2**	Modèle n°3	Modèle n°4
Cohérence	5	0.3586	0.3739	0.3765	0.4103
	10	0.4364	0.4384	0.4224	0.4198
	15	0.4465	0.4305	0.4402	0.4292
	20	0.4284	0.4408	0.4443	0.4411

Tableau 7-Résultats obtenue dans le deuxième cas

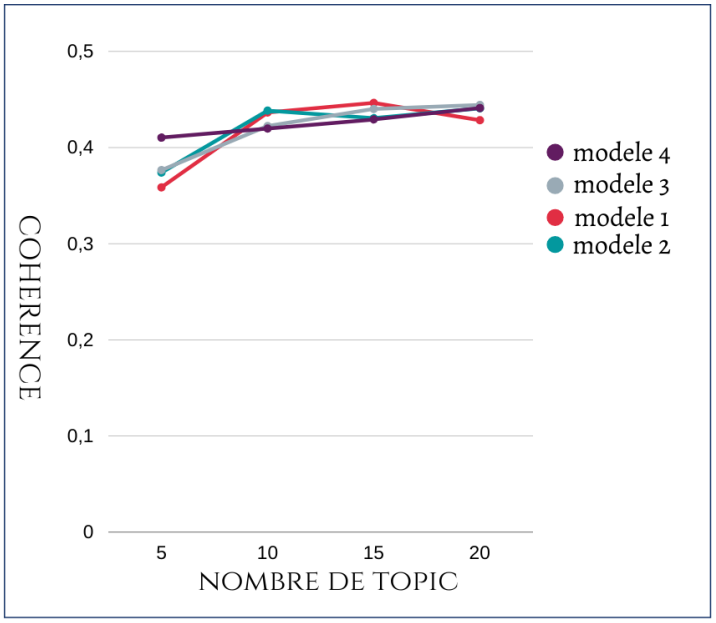

Figure 21-Etude comparative / Cas 2

Partie II

3.II.1 Introduction

Dans cette deuxième partie du troisième chapitre, nous abordons la phase implémentation de notre approche. Nous commençons par la description de l'environnement (plateforme de dèveloppement et d'exécution) et des outils utilisées. En terminant cette partie par un aperçu sur les différentes phases d'exécution du script.

3.II.2 Environnements et technologies

3.II.2.1 Environnements de développement

3.II.2.1.1 Jupyter

Jupyter est un service gratuit pour le développement des logiciels open source, basé sur le web pour les blocs-notes, le code et les données Jupyter. Notre choix est motivé par le fait que les notebooks sont facile à configuré, extensible, modulaire et permet d'organiser l'interface utilisateur pour prendre en charge un large éventail de flux de travail en science des données, en informatique scientifique et en apprentissage automatique. [37]

3.II.2.1.2 python

Python (version 3.7) est un langage de programmation open source le plus populaire, puissant et facile à apprendre. Il a été développé pour la première fois à la fin des années 1980. Il dispose d'une structures de données de haut niveau. Sa syntaxe est élégante,et son typage est dynamique .Ce langage est ideal pour l'écriture de script code et le développement rapide d'applications sur des nombreuses plateformes . Python peut être utilisé pour :

-Rédiger des tests pour vérifier l'état de l'infrastructure informatique

-Ajouter des méthodes aux bibliothèques de manière rétroactive avec singledispatch

- Écrivez à la fois Python et C avec Cython

-Créez des didacticiels interactifs sur la science des données avec Jupyter notebooks [36].

3.II.2.1.3 Canva

Canva est l'outil de conception graphique en ligne le plus populaire au monde. Cet outil permet aux utilisateurs de créer des conceptions, des illustrations, des affiches et des documents et ces créations sont synchronisées avec un compte google, ce qui comprend l'accès de n'importe où et la connexion au programme pour partager et modifier les créations. Il peut être pris en charge dans différents formats (PDF, PNG ...). Canva est une plate-forme simple pour tout le monde où de nombreux éléments sont gratuits mais il existe un certain nombre d'images et de modèles "premium" .[38]

3.II.2.2 Outils utilisés

Nous avons utilisé le package mallet [40] (machine learning for language toolkit), pour exécuter LDA en python sur l'ensemble de données. Mallet est une collection de code basé sur java pour le traitement

du langage naturel, et la classification de documents, modélisation de sujet modélisation et autres applications d'apprentissage automatique pour texte. Les deux entrées principales du modèle LDA mallet sont le **dictionnaire** [id2word] et la **matrice document_terme** (corpus). En plus du dictionnaire et de la matrice, le **nombre de sujets** fait partie également des paramètres essentiels de ce modèle d'apprentissage. Le package mallet utilise l'échantillonnage de Gibbs, pour permettre une implémentation rapide et évolutive de LDA.

.

3.II.3 Implémentation (code + Execution)

Dans cette partie nous décrivons les différentes phases de notre implémentation :

- **Chargement des différentes bibliothèques**

```
import time
import os
import re
import numpy as np
import pandas as pd
from pprint import pprint
import gensim
import gensim.corpora as corpora
from gensim.utils import simple_preprocess
from gensim.models import CoherenceModel
import spacy
import pyLDAvis
import pyLDAvis.gensim
from nltk.corpus import stopwords
stop_words = stopwords.words('english')
```

Figure 22-les differentes bibliothèques

- **Chargement de dataset**

RS_id	researcher	tilte	keywords	abstract
RS_00	A. A. Akintola	A simple and versatile method for frequent 24 ...	Aging;Hormones;Repeated 24 h blood sampling;Sa...	Repeated 24 h blood sampling, which is require...
RS_01	A. A. Ayoola	Analysis of waste groundnut oil biodiesel prod...	ANN;Biodiesel;RSM;Transesterification;Waste gr...	Investigation on the use of KOH and NaOH catal...
RS_02	A. A. Bogush	Changes in composition and lead speciation due...	EfW;Energy-from-Waste;Extraction;Leaching;Muni...	Changes in elemental and mineralogical composi...
RS_03	A. A. Colbourne	Accelerating flow propagator measurements for ...	Dissolution;Interpolation;MRI;NMR;Porous media...	NMR propagator measurements are widely used fo...

Figure 23-Partie du dataset des chercheurs juniors

RS_id	researcher	tilte	keywords	abstract
RS_00	A. Abdelgaied	Comparison of the biomechanical tensile and co...	Compressive biomechanical properties;Decellula...	Meniscal repair is widely used as a treatment ...
RS_00	A. Abdelgaied	A comprehensive combined experimental and comp...	Deep squat;Moderately cross-linked ultra-high ...	A more robust pre-clinical wear simulation fra...
RS_01	A. B. Norton	Poly (vinyl alcohol) modification of low acyl ...	Gellan;Mechanical properties;Phase separation;...	The aim of this research was to control the me...
RS_01	A. B. Norton	Development of 5-(4,6-dichlorotriazinyl) amino...	DTAF;Gellan;Microstructure;Phase separation;St...	Although hydrocolloids are used in a wide rang...
RS_02	A. B. Siegling	Trait emotional intelligence and leadership in...	Cognitive ability;Group differences;Leadership...	This study examined whether trait emotional in...

Figure 24-Partie du dataset des chercheurs seniors

– Prétraitement des données

Après avoir effectué divers techniques de prétraitement du texte, cela comprends: la tokenization, suppression des bruits, suppression des mots d'arrêts, notamment la lemmatization. Nous avons présenté dans la figure ci-contre les résultats obtenus après avoir appliquer ces operations

	données lematisées
doc_0	comparison biomechanical tensile compressive p...
doc_1	comprehensive combine experimental computation...
doc_2	modification low hydrogel application tissue m...
doc_3	development dtaf stain low phase separation st...
doc_4	emotional intelligence leadership multinationa...
...	...
doc_1995	synthesis directly wet air intermediate temper...
doc_1996	conductivity redox stability new oxide redox s...
doc_1997	model predict inhomogeneous protein sugar dist...
doc_1998	adaptable model growth shrinkage droplet respi...
doc_1999	identify critical process step protein stabili...

Figure 25-Résultat aprés le prétraitement de données

– Formation des modèles

Avant d'entraînerr le modèle mallet sur le corpus, il faut mettre à jour le système de variables d'environnement et fournir le chemin vers le fichier Mallet (Figure 26)

```
os.environ.update({'MALLET_HOME':r'C:/mallet-2.0.8'})
mallet_path = 'C:/mallet-2.0.8/bin/mallet'
```

Figure 26-chemin vers le Mallet

Ensuite, comme le montre (la figure 27) nous préparons les deux entrées principales du modèle LDA mallet (dictionnaire et matrice_document_terme)

```
#Préparation des entré de lda model 1
id2word = corpora.Dictionary(data_lemmatized) # Create Dictionary
texts = data_lemmatized # Create Corpus
corpus = [id2word.doc2bow(text) for text in texts] # creat Term Document Frequency
```

Figure 27-préparation des entrées de lda

Enfin arrive l'etape de formation de notre modèle d'apprentissage, en passant le nombre de topic et le chemin du fichier mallet comme paramètre, ainsi que les deux autres paramètre cité précedement.

```
# Train LDA with mallet
ldamallet = gensim.models.wrappers.LdaMallet(mallet_path, corpus=corpus2, num_topics=10, id2word=id2word2)
```

Figure 28- Entraînement du modèle LDA

L'application des quatre modèles d'apprentissage proposés sur les deux dataset (chercheur junior, chercheur senior) nous a permis d'extraire un ensemble de topics et concerne l'affichage des résultat du deuxième modèle LDA

Nombre de topic	Topic	Terme1	Terme2	Terme3
10	Topic 1	49% "effect"	32% "impact"	18% "induce
	Topic 2	43% "method"	39% "study"	31% "simulation"
	Topic 3	50% "base"	33% "assessment"	30% "system"
	Topic 4	70% "cell"	25% "protein"	17% "human"
	Topic 5	39% "flow"	27% "change"	14% "application"
	Topic 6	40% "energy"	24% "cancer"	21% "carbon"
	Topic 7	36% "dynamic"	24% "design"	22% "structure"
	Topic 8	64% "analysis"	18% "rate"	17% "scale"
	Topic 9	81% "modèle"	28% "process"	27% "water"
	Topic 10	35% "surface"	31% "high"	24% "thermal"

15	Topic 1	0.103% "modèle"	78% "base"	32% "process"
	Topic 2	33% "cancer"	20% "health"	13% "clinical"
	Topic 3	58% "energy"	20% "power"	30% "wave"
	Topic 4	20% "organic"	17% "oxide"	15% "population"
	Topic 5	20% "network"	16% "neural"	42% "property"
	Topic 6	40% "water"	17% "mechanical"	16% "chemical"
	Topic 7	19% "natural"	12%"development"	12% "concentration"
	Topic 8	24% "response"	22% "size"	21% "distribution"
	Topic 9	49% "impact"	49% "assessment"	31% "risk"
	Topic 10	97% "analysis"	30% "carbon"	25% "gas"
	Topic 11	54% "study"	51% "dynamic"	44% "system"
	Topic 12	21% "vaccine"	18% "virus"	14% "adaptation"
	Topic 13	0.102% "cell"	23% "human"	17% "memory"
	Topic 14	38% "thermal"	30% "treatment"	23% "temperature"
	Topic 15	35% "protein"	16% "delivery"	18% "food
20	Topic 1	56% "property"	27% "distribution"	27% "size"
	Topic 2	82% "effect"	36% "plant"	35% "gas"
	Topic 3	32% "vaccine"	32% "human"	27% "health"
	Topic 4	54% "protein"	42% "activity"	40% "low"
	Topic 5	91% "high"	56% "process"	29% "detection"
	Topic 6	25% "experimenta	24% "primary"	24% "disease"
	Topic 7	76% "method"	37% "wave"	56% "simulation"
	Topic 8	69% "assessment"	31% "quality"	31% "impact"
	Topic 9	75% "dynamic"	33% "interaction"	30% "network"
	Topic 10	72% "energy"	27% "structural"	24% "power"
	Topic 11	26% "material"	22% "application"	13.4% "analysis
	Topic 12	0.167% "modèle"	33% "factor"	33% "transport"
	Topic 13	53% "change"	40% "carbon"	40% "functional"
	Topic 14	44% "thermal"	35% "treatment"	32% "performance"
	Topic 15	36% "structure"	34% "rate"	33% "patient"
	Topic 16	56% "system"	43% "study "	34% "risk"
	Topic 17	46% "case"	38% "datum"	35% "time"
	Topic 18	0.124% "cell"	39% "cancer"	28% "response"
	Topic 19	76% "surface"	55% "water"	48% "design"
	Topic 20	35% "scale"	32% "spatial"	30% "small"

Tableau 8-Les topic extraits avec le deuxième modèle LDA sur le dataset des chercheurs juniors

Nombre de topic	Topic	Terme1	Terme2	Terme3
12	Topic 1	34%"vaccine"	16%"clinical"	11%"patient"
	Topic 2	84%"energy"	22%"power"	20%"resource"
	Topic 3	43%"water"	36%"flow"	20%"wave"
	Topic 4	75%"analysis"	34%"structure"	28%"dynamic"
	Topic 5	32%"study"	23%"health"	19%"social"
	Topic 6	23%"environmental"	20%"urban"	12%"climate"
	Topic 7	27%"production" +	26%"temperature"	17%"metal"
	Topic 8	23%"transition"	22%"transport"	20%"technology"
	Topic 9	51%"cell"	32%"human"	21%"test"
	Topic 10	46%"risk"	33%"process"	33%"surface"

	Topic 11	33%"carbon"	18%"soil"	11%"organic"
	Topic 12	73%"base"	32%"control"	22%"quality"
15	Topic 1	15.6% "modèle"	29% "modèleling"	27% "simulation
	Topic 2	78%"system"	51%"water"	38% "change"
	Topic 3	40%"thermal"	29%"stress"	28% "induce"
	Topic 4	27%"resource"	26%"material"	23% "growth"
	Topic 5	94%"analysis"	87%"base"	50% "method"
	Topic 6	53%"high"	45%"flow"	29% "heat"
	Topic 7	49%"study"	38%"human"	29% "exposure"
	Topic 8	10.6%"energy"	41%"vaccine"	39% "surface"
	Topic 9	48%"performance"	42%"structure"	23% "soil"
	Topic 10	64%"cell"	37%"response"	34% "low"
	Topic 11	57%"risk"	25%"social"	24% "factor"
	Topic 12	80%"effect"	41%"process"	33% "production"
	Topic 13	30%"design"	29%"transport"	26% "urban"
	Topic 14	41%"control"	38%"food"	24% "consumption"
	Topic 15	65%"assessment"	37%"carbon"	28% "power"
18	Topic 1	12.4% "energy"	90% "system"	29% "technology"
	Topic 2	50% "thermal"	40% "temperature"	33% "induce"
	Topic 3	40% "policy"	31% "urban"	29% "factor"
	Topic 4	35% "heat"	35% "power"	31% "cycle"
	Topic 5	65% "risk	48% "vaccine"	40% "response"
	Topic 6	17.8% "modèle"	10.1% "base"	40% "dynamic"
	Topic 7	49% "surfa	47% "human"	25% "texture"
	Topic 8	80% "cell"	33% "material"	28% "function"
	Topic 9	63% "water"	45% "change"	45% "food"
	Topic 10	11.4% "analysis"	57% "method"	34% "health"
	Topic 11	47% "carbon"	40% "production"	36% "low"
	Topic 12	89% "assessment"	68% "high"	39% "environmental"
	Topic 13	52% "structure"	31% "test"	31% "protein"
	Topic 14	56% "performance"	52% "flow	32% "resource"
	Topic 15	36% "study"	27% "social"	26% "child"
	Topic 16	94% "effect"	27% "rate"	22% "ratio"
	Topic 17	33% "transport"	29% "memory"	25% "source"
	Topic 18	50% "control"	38% "exposure"	34% "image"

Tableau 9-Les topic extraits avec le deuxième modèle LDA sur le dataset des chercheurs seniors

La matrice documents-topics est créée par la suite. Dans le premier cas, en fixant le paramètre num_topic (nombre de sujets) à 15 topic, vu que ce nombre offre les topic les mieux interprétables selon les résultats obtenus (tableau8). Cette même matrice représente également la matrice chercheur-topic (figure 29).

RS_id	Topic_0	Topic_1	Topic_2	Topic_3	Topic_4	Topic_5	Topic_6	Topic_7	Topic_8	Topic_9	Topic_10
RS_00	0.060764	0.052083	0.052083	0.071181	0.069444	0.071181	0.069444	0.052083	0.055556	0.097222	0.053819
RS_01	0.059387	0.057471	0.057471	0.074713	0.072797	0.057471	0.065134	0.074713	0.057471	0.118774	0.057471
RS_02	0.052083	0.052083	0.067708	0.130208	0.067708	0.083333	0.052083	0.065972	0.053819	0.053819	0.076389
RS_03	0.071038	0.056466	0.054645	0.054645	0.056466	0.087432	0.085610	0.071038	0.054645	0.054645	0.129326
RS_04	0.057143	0.209524	0.049206	0.047619	0.090476	0.049206	0.047619	0.050794	0.057143	0.061905	0.050794
RS_05	0.069444	0.059028	0.083333	0.067708	0.083333	0.083333	0.053819	0.088542	0.052083	0.052083	0.065972
RS_06	0.074713	0.057471	0.065134	0.078544	0.057471	0.082375	0.082375	0.078544	0.061303	0.070881	0.059387
RS_07	0.056497	0.060264	0.088512	0.056497	0.056497	0.056497	0.056497	0.056497	0.071563	0.075330	0.056497
RS_08	0.098958	0.052083	0.053819	0.067708	0.062500	0.052083	0.064236	0.083333	0.083333	0.052083	0.052083
RS_09	0.061728	0.061728	0.065844	0.061728	0.061728	0.061728	0.061728	0.063786	0.078189	0.080247	0.080247

Figure 29-Partie de matrice chercheur-topic issue du deuxième modèle LDA /cas1

Contrairement au premier cas, dans le deuxième cas (chercheurs seniors) la matrice document-topic (Figure 30) et la matricechercheur-topic sont tout a fait différentes. Dans la mesure ou nous avons agrégé les topics pour chaque chercheur en utilisant la fonction de maximum (eq.2) (figure 31) ainsi que la fonction de moyenne(eq.1) (Figure 32). En fixant le nombre de sujets à 12.

RS_id	Topic_0	Topic_1	Topic_2	Topic_3	Topic_4	Topic_5	Topic_6	Topic_7	Topic_8	Topic_9	Topic_10
RS_00	0.080247	0.097884	0.145503	0.082011	0.066138	0.067901	0.082011	0.066138	0.066138	0.066138	0.083774
RS_00	0.070776	0.070776	0.121005	0.101218	0.095129	0.072298	0.113394	0.057078	0.096651	0.058600	0.072298
RS_01	0.087571	0.070621	0.070621	0.104520	0.078154	0.070621	0.074388	0.070621	0.070621	0.159134	0.070621
RS_01	0.078947	0.073099	0.084795	0.073099	0.096491	0.090643	0.100390	0.073099	0.073099	0.108187	0.073099
RS_02	0.070128	0.068306	0.068306	0.068306	0.186703	0.068306	0.068306	0.086521	0.082878	0.068306	0.068306
RS_02	0.068306	0.082878	0.068306	0.068306	0.183060	0.068306	0.086521	0.068306	0.082878	0.068306	0.068306
RS_03	0.063847	0.062189	0.080431	0.223051	0.075456	0.062189	0.077114	0.062189	0.077114	0.080431	0.073798
RS_03	0.068996	0.072581	0.067204	0.171147	0.083333	0.067204	0.108423	0.068996	0.070789	0.085125	0.068996
RS_03	0.075980	0.075980	0.061275	0.205065	0.061275	0.061275	0.090686	0.061275	0.062908	0.095588	0.085784
RS_04	0.064103	0.077778	0.229915	0.064103	0.064103	0.079487	0.064103	0.064103	0.069231	0.064103	0.094872
RS_04	0.063131	0.076599	0.226431	0.064815	0.063131	0.073232	0.063131	0.066498	0.078283	0.069865	0.088384
RS_04	0.075980	0.061275	0.149510	0.090686	0.089052	0.075980	0.090686	0.062908	0.061275	0.075980	0.105392
RS_04	0.057870	0.057870	0.198302	0.084105	0.062500	0.060957	0.099537	0.059414	0.068673	0.099537	0.093364
RS_04	0.070621	0.070621	0.155367	0.087571	0.070621	0.070621	0.087571	0.070621	0.070621	0.070621	0.104520

Figure 30-Partie de la matrice docuemnt-topic issue du deuxième modèle LDA / cas2

RS_id	Topic_0	Topic_1	Topic_2	Topic_3	Topic_4	Topic_5	Topic_6	Topic_7	Topic_8	Topic_9	Topic_10
RS_00	0.080247	0.097884	0.145503	0.101218	0.095129	0.072298	0.113394	0.066138	0.096651	0.066138	0.083774
RS_01	0.087571	0.073099	0.084795	0.104520	0.096491	0.090643	0.100390	0.073099	0.073099	0.159134	0.073099
RS_02	0.070128	0.082878	0.068306	0.068306	0.186703	0.068306	0.086521	0.086521	0.082878	0.068306	0.068306
RS_03	0.075980	0.075980	0.080431	0.223051	0.083333	0.067204	0.108423	0.068996	0.077114	0.095588	0.085784
RS_04	0.075980	0.077778	0.229915	0.090686	0.089052	0.079487	0.099537	0.070621	0.078283	0.099537	0.105392

Figure 31-Partie de la matrice chercheur-topic issue de l'agrégation par maximum (eq.2)

RS_id	Topic_0	Topic_1	Topic_2	Topic_3	Topic_4	Topic_5	Topic_6	Topic_7	Topic_8	Topic_9	Topic_10
RS_00	0.076	0.084	0.133	0.092	0.081	0.070	0.098	0.062	0.081	0.062	0.078
RS_01	0.083	0.072	0.078	0.089	0.087	0.081	0.087	0.072	0.072	0.134	0.072
RS_02	0.069	0.076	0.068	0.068	0.185	0.068	0.077	0.077	0.083	0.068	0.068
RS_03	0.070	0.070	0.070	0.200	0.073	0.064	0.092	0.064	0.070	0.087	0.076
RS_04	0.066	0.069	0.192	0.078	0.070	0.072	0.081	0.065	0.070	0.076	0.097

Figure 32-Partie de la matrice chercheur-topic issue de l'agrégation par moyenne (eq.1)

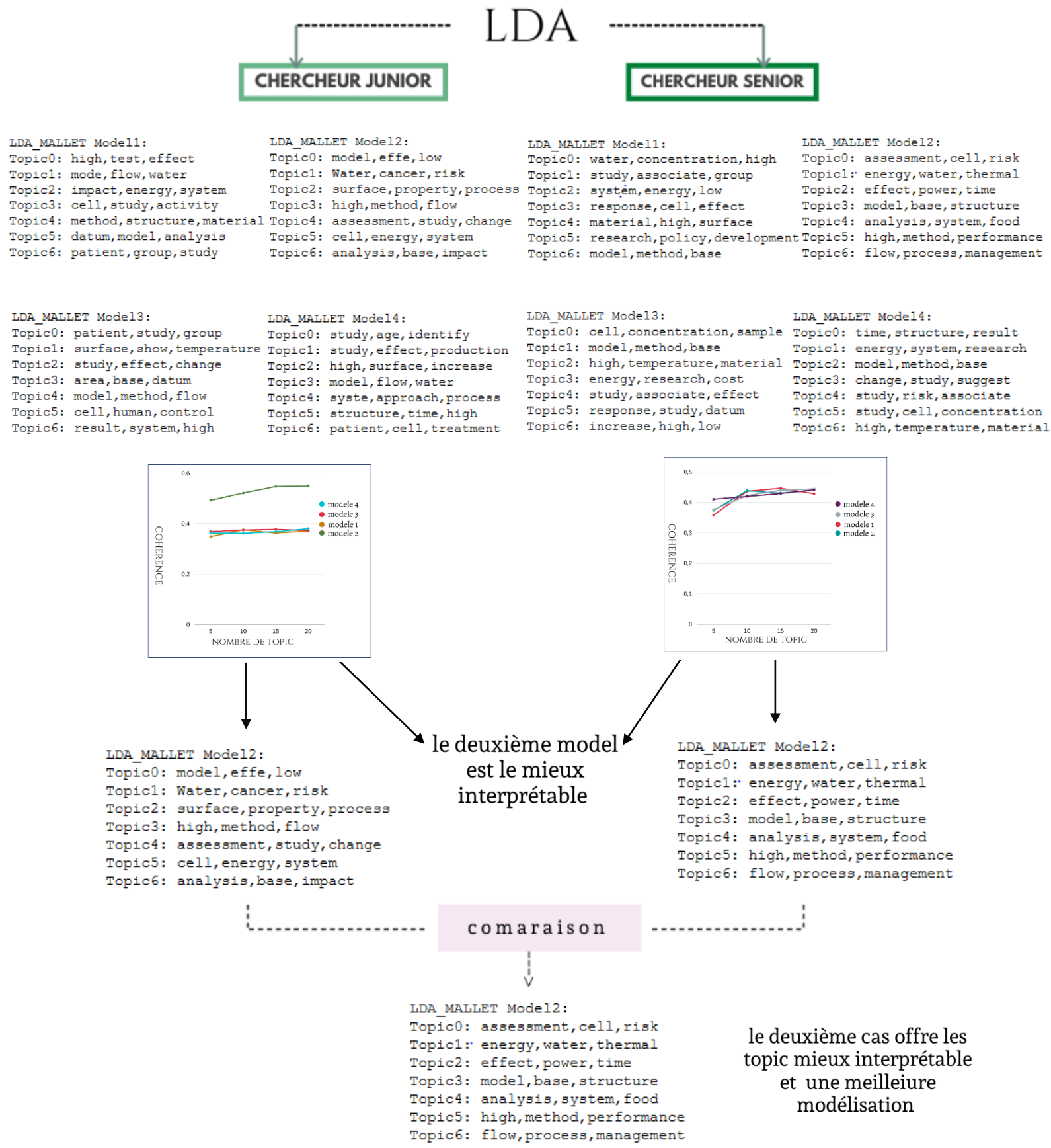

Figure 33-Résumé de l'approche proposée

3.II.4 Discussion des Résultats

À partir des résultats obtenus dans les deux cas de figures. Nous remarquons que le modèle base sur les deux attributs “titre et mots clés” se distingue des autres modèles par la valeur de cohérence maximal, de plus quel que soit le nombre de topics sa valeur de cohérence reste constante, de ce fait nous déduisons finalement que ce modèle est le mieux adapté pour notre jeu de données, et produit également les sujets les mieux précis (compréhensible). Ceci est lié au fait que les deux caractéristiques, titre et mots clés sont choisi par les auteurs de l'article en question, et renvoient par la même l’essentiel des sujets abordés dans ce dernier.

Autre constatation les topics généré dans le cas d’un chercheur ayant dans son corpus plus d’un article, sont composés de termes constituant une structure sémantique précise et compréhensible par rapport au topic produit dans le cas d'un chercheur ayant un seul article dans son corpus.

3.II.5 Exemple illustratif

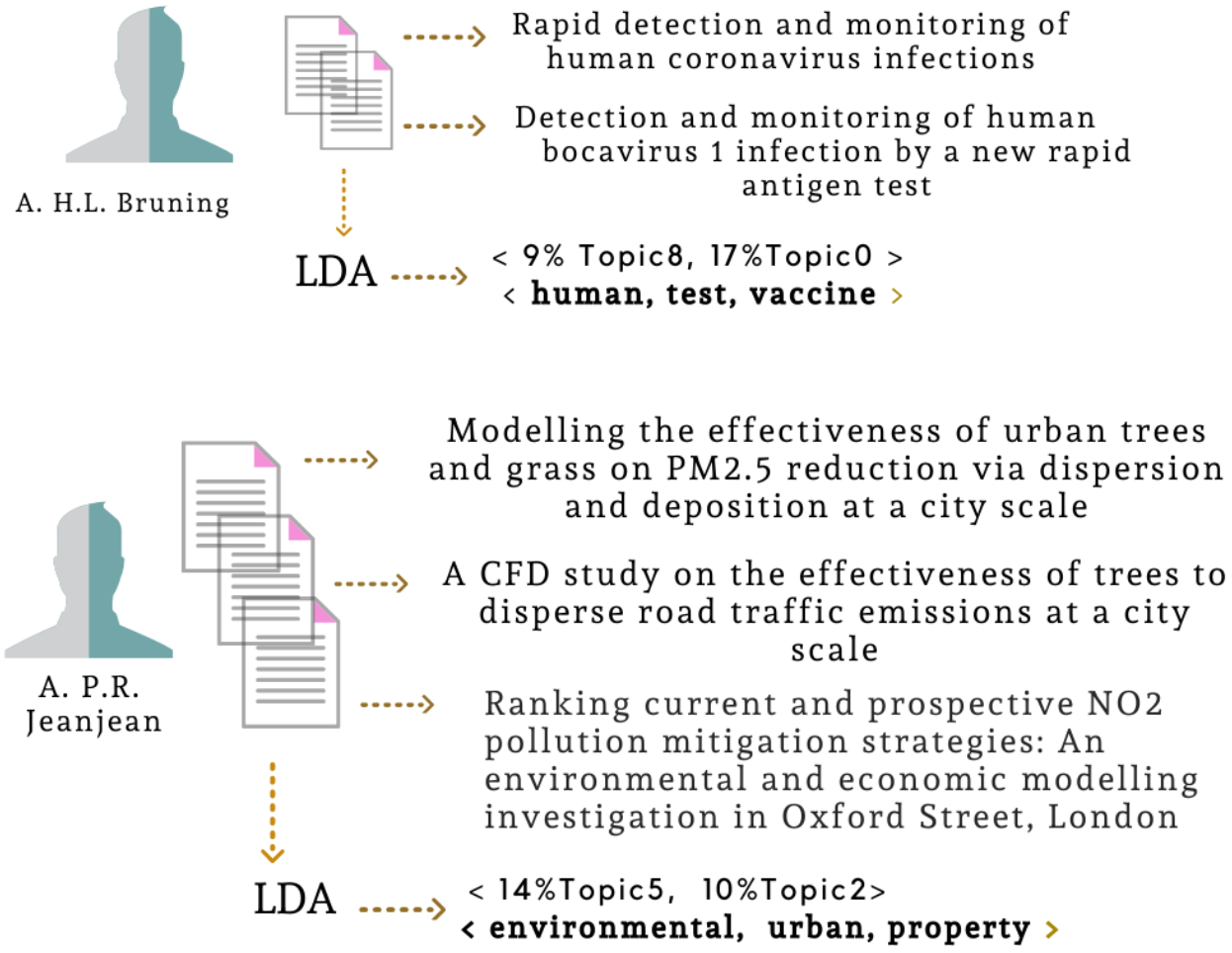

Figure 34-modélisation d'un chercheur senior

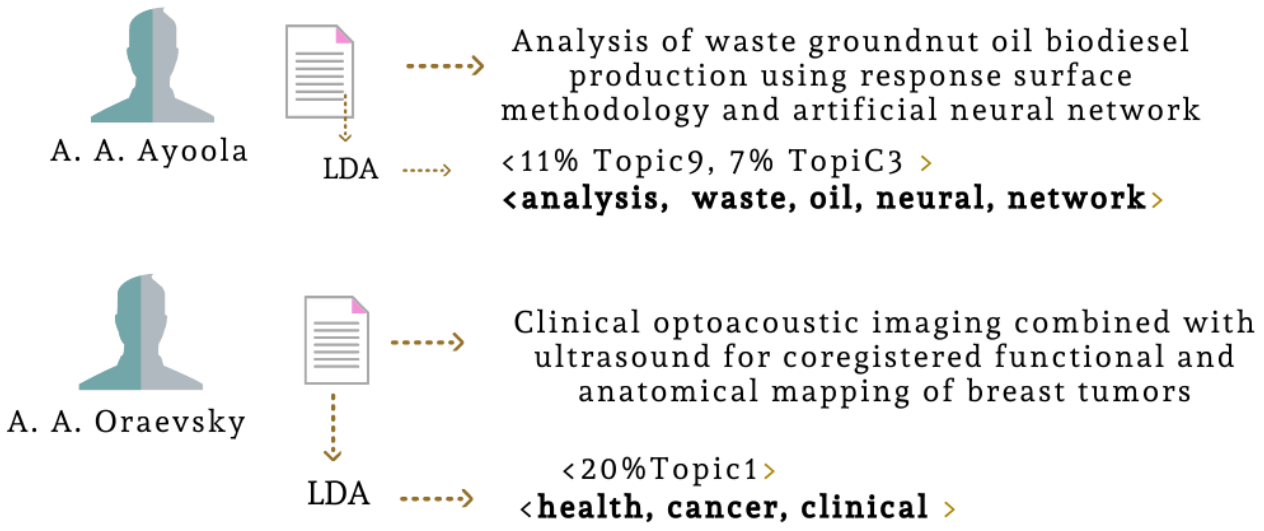

Figure 35-modélisation d'un chercheur junior

3.3 Conclusion

Ce chapitre à fait l'objet de l'approche que nous avons proposée pour modéliser les chercheurs. Nous l'avons divisé en deux parties fondamentales. La première, concerne notre approche avec ses différentes phases. Tandis que la deuxième se concentre principalement sur l'environnement de développement et les divers outils utilisés afin d'implémenter notre approche suivis de quelques parties du code et de l'exécution. En fin par consentement, nous avons conclu que le modèle basé sur la combinaison de caractéristique titre et mot clé fournis les meilleurs résultats, ainsi les topics les mieux compréhensibles.

Conclusion générale

L'analyse des RSNs, est désormais l'un des domaines de recherches les plus développées, c'est pour cela que nous nous sommes intéressés à ce thème. Par caractère cumulatif, nous avons déduit que la modélisation de sujet à l'aide LDA, permet réellement de modéliser efficacement des profils thématiques qui reflète au mieux les domaines d'expertise des chercheurs et leur axe de recherche. Afin d'atteindre nos objectifs, nous avons commencé dans un premier temps, par comprendre les notions de bases des RSAs, ainsi que la structure des publications scientifiques. Ensuite, nous avons présenté notre approche dont le but principal, consiste à représenter chaque chercheur par un vecteur de sujet reflètant ses intérêts.

Perspectives

Etant données que tous mécanismes d'adaptation de l'information (personnalisation, recommandation) s'appuient sur un profil utilisateur. Nous pouvons citer comme perspectives:

- la recommandation des publications récentes liées aux thématiques qui intéressent le chercheur
- L'amélioration des résultats des moteurs de recherche au sein du RSAs
- Le regroupement des chercheurs partageant des intérêts communs

Bibliographie

[1] Elsi Jiménez , Universidad Central de Venezuela . EL USO DE REDES SOCIALES ACADÉMICAS POR LA ACADEMIA . diciembre 2020

[2] https://wearesocial.com/digital-2020

[3] B.Grangerab, S.Tézenas, M.Fagesc, O.Talvardc, T.Similowskid, P.Rufata . Analyse de réseaux sociaux appliquée aux données issus du PMSI national : le parcours du patient BPCO au sein d'AP–HP.6, Revue d'Épidémiologie et de Santé Publique. May 2019

[4] Jooho Kim, Makarand Hastak. Social network analysis: Characteristics of online social networks after a disaster, International Journal of Information Management.2018

[5] Christofer Laurel, Christian Sandström, Adam Berthold, Daniel Larsson.Exploring barriers to adoption of Virtual Reality through Social Media Analytics and Machine Learning An assessment of technology, network, price and trialability, Journal of Business Research. juillet n2019

[6] A. Sayed Mohamed Jelani1, K. Ashkar , R. Sarasu . Research Gate: An Ideal Epitome to Academic Social Networking Sites ; Asian Journal of Information Science and Technology. February 2019

[7] Stefania Manca,ResearchGate and Academia.edu as networked socio-technical systems for scholarly communication: a literature review.25 January 2018

[8] www.researchGtae.com

[9] Shuo Yu Jiaying Liu Feng Xia Xiangjie Kong, Yajie Shi. Academic social networks : Modèleing, analysis, mining and applications. 23 Février 2019

[13] www.academia.edu.com

[14] https://wearesocial.com/fr/blog/2018/03/la-revue-du-social-399-actualite-social-media-facebook-cambridge-analytica

[15] Eamon Costello , Tom Farrelly , and Tony Murphy , Open and Shut : Open Access in Hybrid Educational Technology Journals 2010 – 2017 : International Review of Research in Open and Distributed Learning. January – 2020

[16] Fran¸coise Gouzi. Réseaux sociaux académiques : fonctionnalités principales et en jeux(academia, rasearchgate). 12 mai 2017.

[17] Jacob Eisenstein. . Natural Language Processing . 13 November 2018

[18] Ajit Singh, Natural Language Processing : concept and implémentation with python . 30 mai 2019,117 page.
[19] Frank Millstein. Natural Language Processing with Python: Natural Language Processing Using NLTK : CreateSpace Independent Publishing Platform . 13mars2018, 120 page
[20] Riley Adams,Data Analytics for Businesses 2019: Master Data Science with Optimised Marketing Strategies using Data Mining Algorithms (Artificial Intelligence, Machine Learning, Predictive Modèleling and more) . 5 avril 2019,154 page .
[21] Xu.The Effect of the Multi-Layer Text Summarization Modèle on the Efficiency and Relevancy of the Vector Space-based Information Retrieval.International Journal of Computer Science and Information Security (IJCSIS), 3 mars 2020, Vol. 18 No.
[22] Ezequiel Alvarez, Federico Lamagna, Cesar Miquel Manuel Szewc. Intelligent Arxiv: Sort daily papers by learning users topics preference, février 2020
[23] Elsayed Sabry Abdelaal Issa.the summarization of arabic news texts using probabilistic topic moddeling for L2 micro learning tasks,université arizona.2020
[24] Kai-Xu Han , Wei Chien , Chien-Ching Chiu ,Yu-Ting Cheng .Application of Support Vector Machine (SVM) in the Sentiment Analysis of Twitter DataSet.Applied sciences,2020,vol 10.
[25] Moulana Mohammed, R.M.Noorullahb. Multi Aspects Topic Modèle for Twitter Healthcare Recommendation, 2020, p 5.
[26] Subasish Das,Anandi Dutta,Marcus Brewer.Transportation Research Record Articles: A Case Study of Trend Mining, 14 février 2020.
[27] David Gefen, Jorge E. Fresneda, Kai R. Larsen.Trust and Distrust as Artifacts of Language: A Latent Semantic Approach to Studying Their Linguistic Correlates,26mars2020.
[28] S. Momtazi, A. Rahbar, D. Salami,I. Khanijazani .A Joint Semantic Vector Representation Modèle for Text Clustering and Classification. Journal of AI and Data Mining, 2019 ,Vol 7, No
[29] Oumayma boulaalam,BAdraddine Aghoutane, DRiss EL Ouadghiri, ANiss Moumen ,MOhamed Laghdaf CHeikh Malinine.Design of a tourism recommendation system based on user`s profile.Advanced intelligent system for sustainable development (AI2SD'2019).Morocco:Mostafa Ezziyyani,2019,pages 217-223
[30] Denzler**,** Alexander. User Profiles and Modèles. Granular Knowledge Cube :An Expert Finder System for Knowledge Carriers.2019. Pages 21-35
[31] Christopher Ifeanyi Eke, Azah Anir Norman , Liyana Shuib, Henry Friday Nweke . A Survey of User Profiling: State-of-the-Art, Challenges, and Solutions,vol .7,No. 19037875, 27 September 2019,Pages 144907 – 144924 .

[32] Aarti Singh, Anu Sharma .A Multi-Agent Framework for Contex-Aware Dynamic User oProfiling for web personalization .Software ENgineering PRoceedings of CSI 2015 , India : M.N.Hoda ,S.M.K.quadri,NAresh CHauhan,praven RAnjan Srivastava, 2018 , pages 1-16.

[33] Sirinya ON-AT.Temporalité et réseaux sociaux:prise en compte de l'évolution dans la construction du profil utilisateur. Thèse de doctorat, l'université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier), 29/05/2017.

[34] Vincentas Lamanauskas. SCIENTIFIC ARTICLE PREPARATION: TITLE, ABSTRACT AND KEYWORDS.2019, Vol. 77, No. 4.

[35] M. Rosen-Zvi, T. Griffiths, M. Steyvers, & P. Smyth, 2004. The author-topic modèle for authors and documents. Dans les actes de Proceedings of the 20th conference on Uncertainty in artificial intelligence, 487–494. AUAI Press.

[36] H. M. Wallach, I. Murray, R. Salakhutdinov, & D. Mimno, 2009. Evaluation methods for topic modèles. Dans les actesde Proceedings of the 26th Annual International Conference on Machine Learning, 1105–1112. ACM

[37] T.L.Griffiths&M.Steyvers,2004. Findingscientifictopics. Proceedings of the National academy of Sciences of the United States of America 101(Suppl 1), 5228–5235.

[38] D. M. Blei & J. D. Lafferty, 2009. Topic modèles. Text mining : classification, clustering, and applications 10, 71

[39] Keith Stevens,Philip Kegelmeyer, David Andrzejewski, David Buttler. Exploring Topic Cohérence over many modèles and many topics. Jeju Island, Korea, July 2012, pages 952–961.

[40] http://www. cs.umass.edu

[41] https://opensource.com/resources/python

[42] https://jupyter.org/

[43] https://c-marketing.eu/canva-outil-gratuit-creation-graphique/