République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Chadli Bendjedid
Faculté des Sciences et de la Technologie
Département d'Informatique

الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي والبحث العلمي
جامعة الشاذلي بن جديد
كلية العلوم و التكنولوجيا
قسم الاعلام الالي

جامعة الشاذلي بن جديد
UNIVERSITE CHADLI BENDJEDID

MEMOIRE

Présenté par

GHERIB Amani Malek

Pour l'obtention de diplôme de

MASTER

Filière : Informatique

Spécialité : Systèmes Informatiques Intelligents

SYSTEME COLLABORATIF POUR LA RECOMMANDATION DES PAPIERS DE RECHERCHE

Soutenue le :

Devant le Jury composé de :

Nom et Prénom		Grade		
Dr GASMI	Ibtissem	MCB	UCBET	Président
Dr ANGUEL	Fouzia	MCB	UCBET	Rapporteur
Dr FERROUM	Assia	MCB	UCBET	Examinateur

Année Universitaire : 2019/2020

Remerciements

Avant tout, Qu'elle me soit permis d'exprimer ma profonde reconnaissance à Dr. ANGUEL Fouzia , qui m'a encouragée

dans ce travail, m'a aidée et dirigée dans mes recherches

Mon remerciement s'adresse aussi aux membres du jury qui on accepter d'examiner mon travail

À tous les professeurs du département informatique de l'Université Chadli Bendjedid qui nous ont enseigné au cours de ces années.

.

je tiens à remercier tout particulièrement ma famille.

.

Enfin et avec un très grand plaisir, je remercie tous mes proches et ami(e)s, qui j'ai toujours soutenu et encouragé au cours de l'élaboration de ce mémoire.

Avec plaisir, je remercie tous celles et ceux qui ont participé de près ou de loin à la réalisation de ce travail.

Merci à toutes et à tous.

Dédicace

Avant tout je rends grâce à dieu de m'avoir donné la force et le courage d'achever ce travail, je dédie ce modeste travail à :

Mon père Farid qui a toujours su être là dans tous les moments, les bons comme les difficiles ; son soutien et sa générosité constante ont été fondamentaux pour moi.

Ma mère Ismahane qui a toujours m'écouter et pris le temps d'entendre mes histoires et mes problèmes ; sa douceur, sa tendresse et sa volonté ont toujours mérité mon plus profond respect.

Mes sœurs Ines et Nada et mon frère Khalil qui ont toujours été à mes côtés et m'ont encouragé , Je vous souhaite une vie prospère pleine d'amour et de joie et de succès , que la vie ne puisse jamais nous séparer.

A toute ma famille de près et de loin.

A mon encadreur : « Mm. ANGUEL Fouzia » pour son soutien au moment difficile de mon travail.

GHERIB Amani Malek

Table des matières

Tables des figures

Liste des tableaux

Liste des acronymes

La signification des acronymes utilisés dans ce manuscrit est, en règle générale, précisée lors de leur première utilisation. Ci-après nous donnons tous ces acronymes, leur signification en anglais et (ou) une équivalence en français lorsque nécessaire.

.

TF	Term Frequency.
IDF	Inverse Document Frequency.
SIM	Similarity
FBC	Content-based Filtering
RD	Recherche Documentaire
CF	Collaborative Filtering
SR	système de recommendation
LDA	Allocation de Dirichlet latente
GB	basé sur les graphes
BRG	graphe bi-relationnel
HM	méthodes hybrides (HM)
CP	papiers candidats

Introduction Générale

Généralement, sur les plateformes en ligne les utilisateurs peuvent générer et partager du contenu. Une variété d'items allant des produits proposés dans une boutique en ligne aux publications sur les réseaux sociaux peuvent être évalués ou notés par les utilisateurs. Par exemple, sur la plateforme de distribution d'œuvres cinématographiques et télévisuelles Netflix[1], un utilisateur peut fournir des commentaires d'un simple clic de la souris en utilisant le système de classement à cinq étoiles qui spécifient ses goûts ou ses aversions d'un article. Dans d'autres sites Web comme Amazon[2], la modification des détails d'un produit peut être considérée comme une note positive implicite pour ce produit. Ces formes d'informations devraient être gérées et traitées pour créer des applications qui adaptent leurs fonctionnalités aux besoins des utilisateurs.

En réponse à ce problème, des systèmes de recommandation ont été développé. Les systèmes de recommandations sont des applications logicielles qui se basent sur les intérêts des utilisateurs, ses notes et son comportement précédent, pour suggérer des éléments que les utilisateurs sont susceptibles de préférer parmi une énorme collection d'items. Ces systèmes utilisent des techniques de filtrage pour fournir des recommandations.

Les systèmes de recommandation ont été appliqués avec succès dans plusieurs domaines, suggérant des items tels que des films, de la musique, des livres, des produits et également les articles scientifiques.

Dans le domaine de recherche scientifique, comme la recherche est un processus continu. Alors que de plus en plus d'articles de revues et de documents de conférence sont publiés d'année en année, il devient de plus en plus difficile d'identifier les articles de recherche liés à un domaine d'intérêt. De plus, il devient non trivial de se tenir au courant des nouvelles publications ainsi que de les associer à des articles déjà publiés. En fait, la quantité volumineuse de ces informations rend le processus de recherche d'informations difficile et très ennuyeux. La tâche de retrouver des articles pertinents à partir d'un volume aussi énorme n'est pas évidente, car le système de recherche doit fournir les meilleurs résultats en traitant le Big Data. Le problème s'aggrave lorsque les chercheurs débutants ne peuvent pas trouver leurs articles pertinents en raison du manque d'expérience dans l'utilisation de ces moteurs de recherche. Le processus de filtrage des articles pertinents manuellement est également une tâche fastidieuse et longue. Par conséquent,

[1] www.netflix.com

[2] www.amazon.fr

un système de recommandation efficace des articles de recherche est nécessaire pour produire des recommandations de haute qualité à partir de ces entrepôts numériques.

Dans cette issue, les systèmes de recommandation sont utilisés pour la recherche d'articles scientifiques pertinents au domaine d'intérêt d'un chercheur. Ainsi, Un système de recommandation peut suggérer automatiquement des articles scientifiques basés sur les préférences de l'utilisateur ou / et d'autres utilisateurs ayant des intérêts similaires.

Dans les dernières années, de nombreux systèmes de recommandation des articles scientifiques ont été mis en œuvre reposant sur différentes méthodes. Ces méthodes consistent en un filtrage collaboratif, filtrage basé sur le contenu ou des techniques basées sur l'analyse des citations. C'est sur cet aspect que nous nous focaliserons dans ce mémoire. Notons que dans ce mémoire nous utilisons indifféremment les termes article scientifique et papier de recherche.

1. Objectifs

L'objectif de ce mémoire est d'étudier le problème de recommandation des articles de recherche Dans ce travail nous présentons une approche collaborative pour la recommandation des articles scientifiques basé sur le réseau des citations.

2. Contenu du mémoire

Ce mémoire est organisé comme suit : une introduction générale, trois chapitres, et une conclusion avec des perspectives.

- **Chapitre 1 (état de l'art) :** offre un survol des systèmes de recommandation. Dans ce survol nous discutons les différentes approches connues dans ce domaine tout en mettant en relief leurs utilisations pour la recommandation d'articles de recherche.
- **Chapitre 2 (conception) :** ce chapitre présente la solution proposée tout en détaillons ses fonctionnalités.
- **Chapitre 3 (implémentation) :** après avoir présenté les aspects techniques entourant l'implémentation de l'approche proposée ainsi que la description des différentes interfaces de l'application, nous décrivons le détail de l'expérimentation de notre application et nous présentons les résultats de validation de notre système de recommandation des articles scientifiques de recherche.
- Finalement, nous terminons ce mémoire par une conclusion et nous présentons quelques perspectives futures.

"Courage, espoir et confiance en soi sont les prémisses puissantes d'un fondement véhiculaire de la voie de la réussite."

Mofaddel Abderrahim

Chapitre 1 : Etat de l'Art

1. Introduction

Dans le contexte numérique actuel, caractérisé par une surabondance d'informations, il apparaît que les capacités humaines ne permettent pas l'analyse exhaustive de l'offre d'un corpus au sein d'une plateforme. Même dans le cadre de l'utilisation d'un moteur de recherche intégré, les résultats pertinents sont généralement noyés dans un « bruit » informationnel, ce qui en empêche, ou tout du moins en ralentit, le repérage.

Pour aider l'esprit humain, les systèmes de recommandation (SR) ont été introduits comme une technique intelligente pour faire le filtrage de l'information afin de présenter les éléments d'information qui sont susceptibles d'intéresser l'utilisateur. De manière générale, La raison pour laquelle les gens pourraient être intéressés à utiliser un système de recommandation est qu'ils ont tant d'éléments à choisir dans une période limitée de temps et ils ne peuvent pas évaluer toutes les options possibles.

Les systèmes de recommandation peuvent être utilisés pour fournir efficacement des services personnalisés. Par exemple dans le domaine de commerce électronique, les systèmes de recommandation sont bénéfiques pour le client en lui faisant des suggestions sur les produits susceptibles d'être appréciés. En même temps, l'entreprise ou le vendeur va bénéficier de l'augmentation des ventes qui se produit normalement quand on présente au client plus d'articles susceptibles d'être aimés [Vozalis *et al.*, 2003].

Ce chapitre est un survol des systèmes de recommandation. Ce survol présente la définition des systèmes de recommandation, suivie de la présentation de plusieurs logiques de classification de ces systèmes, bien connues dans ce domaine. Ensuite, nous exposons les approches de recommandation des papiers scientifiques. Enfin de ce chapitre nous présentons les différentes méthodes et métriques d'évaluation des systèmes de recommandation.

2. Principes des systèmes de recommandation

2.1 Définition

Les systèmes de recommandation ou (en Anglais recommander systems) ont été définis de plusieurs façons. La définition la plus populaire et la plus générale est celle de Robin Burke [Burke, 2002] que nous avons traduite ainsi : « Un système de recommandation est un système capable de fournir des recommandations personnalisées ou permettant de guider l'utilisateur de manière personnalisée vers des objets intéressants ou utiles au sein d'un espace de données important ».

L'objectif d'un système de recommandation est de fournir à l'utilisateur des objets pertinents selon ses préférences. Il permet de réduire de manière considérable le temps que l'utilisateur met pour chercher les objets les plus intéressants pour lui, et aussi de trouver des objets qu'il est susceptible d'aimer mais auxquels il n'aurait pas forcément fait attention.

Un système de recommandation est un système de filtrage d'informations de façon personnalisée pour chaque utilisateur. Autrement dit, dans un but de personnaliser la recherche d'information dans un domaine d'application particulier, un système de filtrage collecte, sélectionne, classifie et suggère à l'utilisateur les informations qui répondent vraisemblablement à ses intérêts à long termes.

2.2 Eléments d'un système de recommandation

Les deux entités de base qui apparaissent dans tous les systèmes de recommandations sont l'utilisateur et l'item ou article. L'« utilisateur» est la personne qui utilise un système de recommandation, donne son opinion sur diverses items et reçoit les nouvelles recommandations du système.

L'« item » est le terme général utilisé pour désigner ce que le système recommande aux usagers tels qu'un film ou un article scientifique.

Les données d'entrée pour un système de recommandation dépendent du type de l'algorithme de filtrage employé. Généralement, elles appartiennent à l'une des catégories suivantes :

- Les estimations (également appelées les votes ou les notes), qui expriment l'opinion des utilisateurs sur les items. Elles sont normalement fournies de façon explicite par l'utilisateur et suivent une échelle numérique spécifique (exemple : 1 mauvais à 5 excellent). Les estimations peuvent également être collectées implicitement à partir de l'histoire d'achat de l'usager, des logs Web, des habitudes de lecture et d'écoute,…etc;
- Les données démographiques, qui se réfèrent à des informations telles que l'âge, le sexe et l'éducation des utilisateurs. Ce type de données est généralement difficile à obtenir et est normalement collecté explicitement;
- Les données de contenu, qui sont fondées sur une analyse textuelle des documents liés aux éléments évalués par l'utilisateur. Les caractéristiques extraites de cette analyse sont utilisées comme entrées dans l'algorithme de filtrage afin d'en déduire un profil d'utilisateur [Vozalis *et al*., 2003].

3. Classification des systèmes de recommandation

Depuis l'émergence du domaine des systèmes de recommandation, un nombre important de travaux de recherche ont traité la problématique de la recommandation au cours des dernières années. Ces travaux sont issus de plusieurs domaines comme le Machine Learning, les statistiques et surtout la recherche d'information. Plusieurs approches et méthodes ont été proposées, certaines ont été étudiées, expérimentées et comparées.

Parfois plusieurs terminologies sont utilisées pour désigner une même méthode ou approche. C'est pour cela des travaux se sont intéressés à la classification de ces méthodes et proposent une taxonomie ou terminologie unifiée [Adomavicius & Tuzhilin 2005; Burke, 2002].

Il existe plusieurs classifications des systèmes de recommandations (Figure 1) :

- La classification classique des systèmes de recommandation est reconnue par trois types de filtrage ; le filtrage collaboratif(CF), le filtrage basé sur le contenu (CBF) et le filtrage hybride.
- La Classification de Rao et Talwer : elle est utilisée dans les systèmes de filtrage collaboratif. Ils proposent une sous classification qui comprend trois catégories [Rao et Talwer, 2008] :
  - Approches CF a base de mémoire : pour K-plus proches voisins.
  - Approches CF base sur un modèle englobant une variété de techniques telles que: clustering, les réseaux bayesiens, factorisation de matrices, les processus de décision de Markov.
  - Approches CF hybride qui combine une technique de recommandation CF avec une ou plusieurs autres méthodes.

- La classification de Su et al : c'est une classification en fonction de la source d'information utilisée.
- La classification la plus utilisée est une classification selon deux approches : les recommandations basées sur le contenu et le filtrage collaboratif [Burke, 2002]. En plus de ces deux approches, Burke propose de considérer trois autres approches : la recommandation basée sur les données démographiques, la recommandation basée sur la connaissance (knowledge-based) et la recommandation basée sur l'utilité (utility-based). Mais il note que ces trois approches sont des cas particuliers des approches classiques [Burke, 2007].

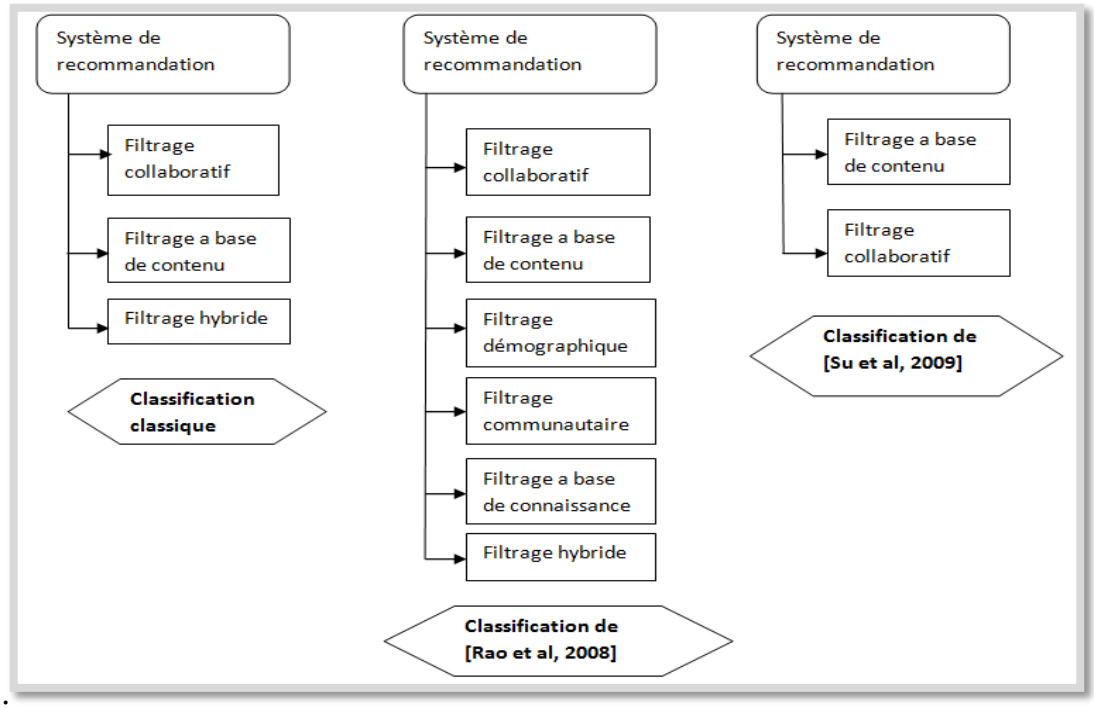

Figure 1 : Classification des systèmes de recommandation.

3.1 Recommandation basée sur le contenu

Les systèmes de recommandations basés sur le contenu (voir figure 2) fournissent des recommandations en comparant la représentation du contenu décrivant un item ou un produit à la représentation du contenu décrivant l'intérêt de l'utilisateur (***profil d'intérêt de l'utilisateur***). Ils sont parfois appelés des systèmes de filtrage basé sur le contenu (FBC : Content-based Filtering), qui est une évolution générale des études sur le filtrage d'information. Le filtrage basé sur le contenu s'appuie sur le contenu des documents (thèmes abordés) pour les comparer à un profil lui-même constitué de thèmes. Chaque utilisateur du système possède alors un profil qui décrit ses points d'intérêts.

Par exemple, le profil peut contenir une liste des thèmes ou préférences que l'utilisateur aime bien ou qu'il n'aime pas. Lors de l'arrivée d'un nouveau document (item), le système compare le descriptif du document (item) avec le profil de l'utilisateur pour prédire l'utilité de ce document pour cet utilisateur.

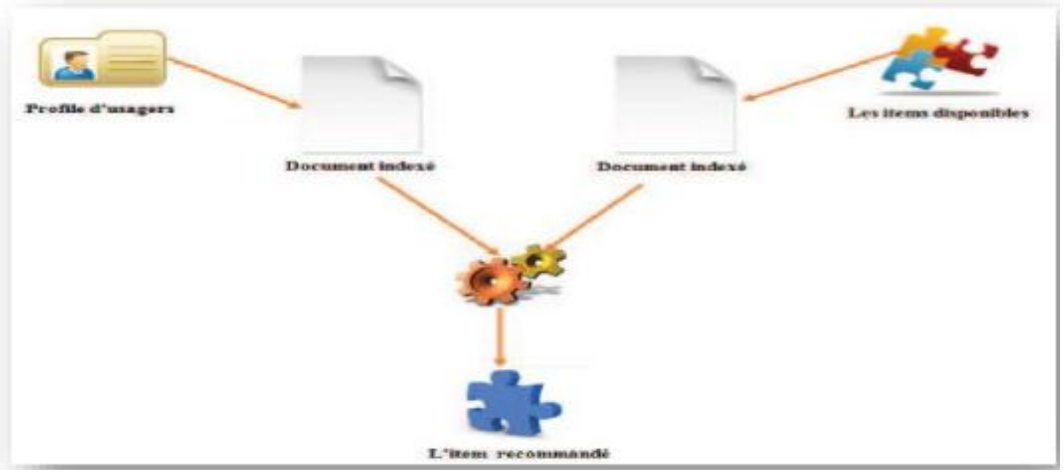

Figure 2: Principe de filtrage à base de contenu. [Lops et al., 2011]

Un système de recommandation basé sur le contenu a besoin de techniques pour produire une représentation efficace des items et du profil de l'utilisateur pour pouvoir les comparer. Ainsi une architecture de haut niveau (figure 3) est proposée dans [Lops et al., 2011] dans laquelle le processus de recommandation est réalisé en trois étapes, chacune étant gérée par un composant spécifique :

- **Analyseur du contenu** : Lorsque l'information n'est pas structurée (par exemple, un item représenté par un texte), ce module a pour but d'en réaliser le prétraitement pour extraire l'information pertinente, la structurer et la représenter dans une forme cible appropriée (par exemple un vecteur de mots clés).
- **Apprentissage du profil** : Ce module collecte les données représentatives des préférences de l'utilisateur et généralise ces données, afin d'apprendre et de construire le profil de l'utilisateur. Des techniques d'apprentissage automatique peuvent être utilisées pour cela. On peut citer à titre d'exemple les arbres de décisions, les réseaux de neurones et la classification naïve de Bayes. Ces techniques visent à inférer un profil de l'utilisateur en utilisant l'information sur les items qu'il a aimés ou n'a pas aimés.

  Afin de construire et mettre à jour le profil de l'utilisateur actif, ses réactions aux items (notes) sont recueillies et enregistrées dans le composant Feedback. Ces notes d'intérêt sont exploitées au cours du processus d'apprentissage du modèle utile pour prédire la pertinence a priori d'un item que l'utilisateur n'a pas encore noté. Les utilisateurs peuvent aussi définir explicitement leurs domaines d'intérêt au préalable comme profil initial, mais ce cas est assez rare.
- **Composant de filtrage** : Ce module filtre les items pertinents en faisant correspondre la représentation du profil utilisateur aux items candidats à la recommandation. La pertinence

de l'item est calculée en utilisant des métriques de similarité entre l'item considéré et le profil de l'utilisateur. Plus la similarité avec le profil "positif" est grande et plus la similarité avec le profil "négatif" est petite, plus l'item a des chances d'être recommandé.

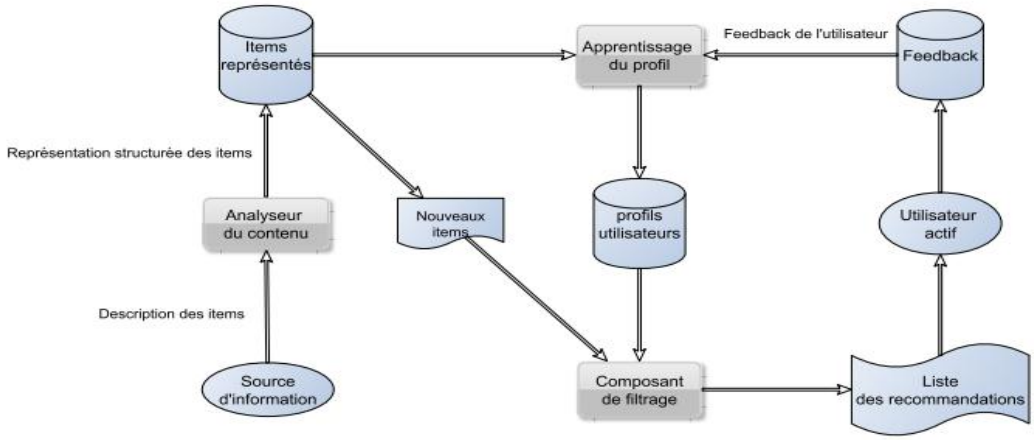

Figure 3 : Une architecture haut niveau d'un système de recommandation basé sur le contenu (d'après [Lops et al., 2011]).

Les approches FBC initiales avaient leurs racines dans le champ de Recherche Documentaire (RD). Les premiers systèmes sont concentrés sur le domaine de type textuel, comme les articles de journaux, les pages Web et ils ont appliqué des techniques de RD pour extraire les informations significatives du texte. Récemment sont apparues quelques solutions qui s'occupent des domaines plus complexes [Arnautu, 2012].

L'élément clé de l'approche FBC est le calcul de similarité entre les items qui indique une distance entre les éléments. Quelques fonctions de distances communes, étant donné deux vecteurs de caractéristique x et y, sont utilisées : la distance euclidienne, la distance Manhattan, la distance Tchebychev, la distance de cosinus pour les vecteurs et la distance Mahalanobis [Arnautu, 2012].

3.1.1 Avantages des systèmes de recommandation à base de contenu

- **Autonomie de l'utilisateur** : les techniques de recommandation basées sur le contenu traitent chaque utilisateur de façon indépendante. Ainsi, seules les évaluations de l'utilisateur lui-même sont prises en compte pour construire son profil utilisateur et faire la recommandation, ce qui n'est pas le cas pour les approches utilisant le filtrage collaboratif [Lops et al., 2011].

- **Nouvel item** : Le filtrage basé sur le contenu peut recommander des items nouvellement introduits dans la base avant même qu'ils reçoivent une évaluation de la part d'un utilisateur, au contraire des approches collaboratives qui ne peuvent recommander un item que s'il a été préalablement évalué par un groupe d'utilisateurs [Ait Ahmed, 2017]

- **Transparence** : Des explications sur le fonctionnement du système de recommandation peuvent être fourni en donnant explicitement le contenu ou les descriptions qui ont causé un élément à être recommandé. Ces fonctionnalités sont des indicateurs à consulter afin de décider de faire confiance à une recommandation. À l'inverse, les systèmes collaboratifs sont des boîtes noires puisque la seule explication d'une recommandation d'un item est que des utilisateurs inconnus ayant des préférences similaires ont aimé cet item [Lops et al., 2011]..

- Le caractère dynamique de ces systèmes est également un avantage car plus l'utilisateur se servira du système et plus la pertinence des items qui lui seront proposés sera fine. En revanche, un utilisateur ne se verra jamais proposer d'items qui n'auront pas été jugés similaires à ceux qu'il a apprécié [Bechet, 2012].

3.1.2 Les inconvénients des systèmes de recommandation à base de contenu

- Limite de l'analyse du contenu : une limite naturelle de la recommandation basée sur le contenu est la nécessité de disposer d'une représentation variée et riche du contenu des items, ce qui n'est pas toujours le cas. La précision des recommandations est liée à la quantité d'informations dont dispose le système pour discriminer les items appréciés de ceux non appréciés par l'utilisateur. Contrairement au filtrage collaboratif qui peut traiter tout type d'items sans aucune information sur leur contenu, l'approche basée sur le contenu ne peut traiter que les items disposant d'un contenu pouvant être analysé.
- Sur-spécialisation (Over-specialization) : le système ne peut recommander que les items qui sont similaires au profil utilisateur. L'utilisateur ne peut donc recevoir que des recommandations proches des items qu'il a notés ou observés par le passé. Or, la diversité des recommandations est souvent appréciée et s'avère être un critère d'évaluation important des systèmes de recommandation. Idéalement, l'utilisateur doit recevoir des recommandations pertinentes et diversifiées.
- Intégration d'un nouvel utilisateur est non immédiate : un utilisateur doit évaluer un certain nombre d'items avant que le système ne puisse interpréter ses préférences et lui fournir des recommandations pertinentes, Ce problème est connu dans la littérature sous le nom du problème de démarrage à froid pour les utilisateurs (user cold start problem) [Naak, 2009].

- Les difficultés à recommander des documents multimédia (images, vidéos, etc.) et ceci à cause de la difficulté à indexer ce type de documents, c'est en fait la même problématique dont souffrent les systèmes de recherche.

3.2 Recommandation basé sur le filtrage collaboratif

Le filtrage collaboratif (Collaborative Filtering « CF ») a pour principe d'exploiter les évaluations faites par des utilisateurs sur certains documents (items), afin de recommander ces mêmes items à d'autres utilisateurs (voir figure 4), et sans qu'il soit nécessaire d'analyser le contenu des documents. Tous les utilisateurs du système de filtrage collaboratif peuvent tirer profit des évaluations des autres en recevant des recommandations pour lesquelles les utilisateurs les plus proches ont émis une évaluation de valeur favorable, et cela sans que le système dispose d'un processus d'extraction du contenu des documents ou extraction des caractéristiques des items. Grace à son indépendance vis-à-vis de la représentation des données, cette technique peut s'appliquer dans les contextes où le contenu est soit indisponible, soit difficile à analyser, et en particulier elle peut s'utiliser pour tout type de données : texte, image, audio et vidéo. De plus, l'utilisateur est capable de découvrir divers domaines intéressants, car le principe du filtrage collaboratif ne se fonde absolument pas sur la dimension thématique des profils.

Le premier système qui a utilisé la méthode de filtrage collaboratif a été le projet Tapestry au Xerox PARC [Goldberg *et al.,* 1992]. Le projet a inventé le terme de filtrage collaboratif. D'autres systèmes pionniers sont : le recommandeur de musique nommé Ringo [Shardanand & Maes, 1995], et Group Lens, un système pour évaluer les articles USENET [Resnick *et al.,* 1994].

Les deux facteurs clés d'un système de filtrage collaboratif sont les profils des utilisateurs et les communautés.

- Le profil utilisateur est composé de prédicats pondérés. Le poids d'un prédicat exprime son intérêt relatif pour l'utilisateur. Il est spécifié par un nombre réel compris entre 0 et 1. Le profil s'enrichit progressivement au fur et à mesure que l'utilisateur évalue les items reçus. Outre les informations d'identification de base.
- La notion de communauté dans un système de filtrage collaboratif est définie comme le regroupement des utilisateurs en fonction de l'historique de leurs évaluations, afin que le système calcule des recommandations.

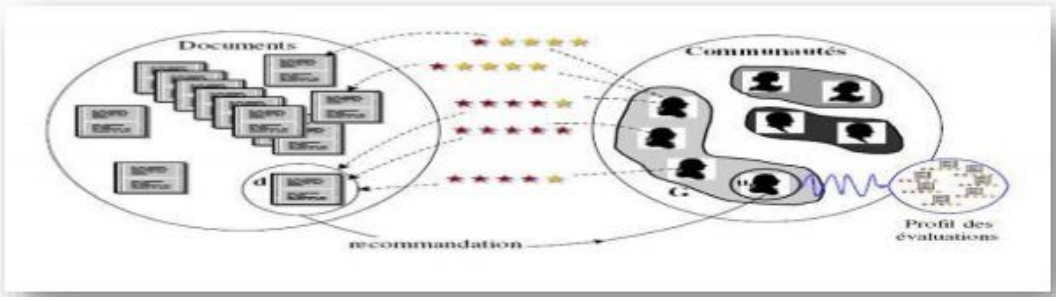

Figure 4: Principe général du filtrage collaboratif.

Le filtrage collaboratif est basé sur un processus qui suit les étapes décrites ci-dessous:

- **Evaluation des recommandations :** Selon le principe de base du filtrage collaboratif, les utilisateurs doivent fournir leurs évaluations sur des items afin que le système forme les communautés. Evaluer une recommandation peut se faire de façon explicite ou implicite, comme suit :
    - Explicite : L'utilisateur donne une valeur numérique sur une échelle donné (par exemple de 1 à 5, ou de 1 à 10, etc.), ou bien, une valeur qualitative de satisfaction, par exemple, mauvaise, moyenne, bonne et excellente.
    - Implicite : Le système induit la satisfaction de l'utilisateur à travers ses actions. Par exemple, le système estimera que la suppression d'un document recommandé correspond à une évaluation très mauvaise, alors que l'impression ou la sauvegarde d'un document recommandé peut être interprétée comme une bonne évaluation.
- **Formation des communautés :** Le processus de formation des communautés est le noyau d'un système de filtrage collaboratif. Pour chaque utilisateur, le système doit calculer sa communauté, généralement cela se fait par la proximité des évaluations des utilisateurs. Pour ce faire, on peut calculer, dans un premier temps, la proximité entre un utilisateur donné et tous les autres. Ensuite, et afin de créer contrairement la communauté de l'utilisateur, on applique la méthode des voisins les plus proche en utilisant un seuil pour le niveau de proximité ou un seuil pour la taille maximale de la communauté, en raison de sa performance et sa précision.
- **Production des recommandations :** Dans ce derniers processus, une fois la communauté de l'utilisateur créée, le système prédit l'intérêt qu'un item particulier peut présenter pour l'utilisateur en s'appuyant sur les évaluations que les membres de la communauté ont fait sur ce même item. Lorsque l'intérêt prédit dépasse un certain seuil, le système recommande l'item à l'utilisateur.

3.2.1 Avantages des systèmes de recommandation basé sur le filtrage collaboratif

- Effet de surprise : l'effet de surprise que peut recevoir l'utilisateur en recevant une recommandation pertinente qu'il n'aurait pas trouvée seul est souvent souhaitable. Les algorithmes basés sur le filtrage collaboratif permettent généralement de faire des recommandations à effet de surprise. Par exemple, si un utilisateur "X" est proche d'un utilisateur "Y" du fait qu'il ne regarde que des comédies, et si "Y" apprécie un film d'un autre genre, ce film peut être recommandé à "X" du fait de sa proximité avec "Y".

- Non nécessitée de la connaissance du domaine : les systèmes de recommandation basés sur le filtrage collaboratif ne requièrent aucune connaissance sur les items. Ces méthodes peuvent recommander des items sans avoir besoin de comprendre leurs sens ni disposer de leurs attributs. La recommandation est basée uniquement sur les notes données aux items [Naak, 2009].

- l'utilisateur est capable de découvrir divers domaines intéressants, car le principe du filtrage collaboratif ne se fonde absolument pas sur la dimension thématique des profils, et n'est pas soumis à l'effet « entonnoir ».

- les évaluations faites par les utilisateurs intègrent non seulement la dimension thématique mais aussi d'autres facteurs relatifs à la qualité des items tels que la diversité, la nouveauté, etc.

3.2.2 Inconvénients des systèmes de recommandation basé sur le filtrage collaboratif

- Le démarrage à froid : concerne à la fois les nouveaux utilisateurs et les nouveaux items qui sont introduits dans le système. Un nouvel utilisateur qui n'a noté aucun item ne peut pas recevoir de recommandation puisque le système ne connaît pas ses préférences. Ce problème est connu sous le nom de problème du démarrage à froid pour les utilisateurs (user cold start).

- La parcimonie (sparsity) : Le nombre d'items candidats à la recommandation est souvent énorme et les utilisateurs ne notent qu'un petit sous-ensemble des items disponibles. De ce fait, la matrice des notes est une matrice creuse avec un taux de valeurs manquantes pouvant atteindre 95% du total des valeurs [Papagelis, 2005]. Les systèmes de filtrage collaboratif ont des difficultés dans ce cas, le nombre de notes à prédire étant largement supérieur aux nombres de notes déjà connues. Le problème de la parcimonie peut être réduit en utilisant les approches par modèles qui réduisent la dimension de la matrice des notes.

- Le problème du mouton gris (gray sheep) : Les utilisateurs qui ont des goûts étranges (qui varient de la norme ou qui sortent du commun) n'auront pas beaucoup d'utilisateurs voisins. Il sera donc difficile de faire des recommandations pertinentes pour ce genre d'utilisateurs.

3.3 Filtrage hybride

Un système de recommandation est dit hybride quand il combine deux ou plusieurs approches de recommandation différentes.

Les approches basées sur le contenu ont l'avantage de pouvoir recommander les nouveaux items non encore évalués par un utilisateur, alors que le filtrage collaboratif ne peut recommander un item que s'il a été noté par un certain nombre d'utilisateurs auparavant. Les approches basées sur le contenu nécessitent de disposer des attributs des items, en plus d'une étape d'analyse pour pouvoir les extraire et les représenter, alors que le filtrage collaboratif ne requiert pas d'accès au contenu des items pour pouvoir faire de la recommandation. Il s'appuie uniquement sur la matrice des notes d'utilisateurs pour les différents items.

L'hybridation de ces deux techniques, afin de traiter les insuffisances de chaque technique utilisée seule et profiter de leurs points forts, a fait l'objet de plusieurs travaux de recherche. Le système Fab [Balabanovic and Shoham, 1997] est un des premiers systèmes de recommandation hybrides. Il combine le filtrage collaboratif et une approche basée sur le contenu afin de traiter à la fois le problème du démarrage à froid pour les items et la sur-spécialisation. Dans ce système, deux critères doivent être satisfaits pour recommander un item : son contenu doit être similaire au profil de l'utilisateur, et il doit être apprécié par les voisins les plus proches.

La combinaison de la méthode FBC et la méthode FC pour construire un système hybride se fait de différentes manières. Selon Burke on peut distinguer sept façons de combiner les méthodes traditionnelles [Burke, 2002] :

- Pondération (Weighted) : le score ou la prédiction obtenu par chacune des deux techniques recommandations est combiné en un seul résultat [Naak, 2009].
- Par sélection (Switching) : le système bascule entre les deux techniques de recommandation en fonction de la situation [Naak, 2009]
- Technique mixte (Mixed) : dans cette approche, le recommandeur ne combine pas, mais augmente la description des ensembles de données, en prenant en considération les estimations des utilisateurs et la description des items. La nouvelle fonction de prédiction

doit faire face aux deux types de descriptions et permet d'éviter les problèmes posés par le filtrage collaboratif, à savoir, le démarrage à froid.

- Par combinaison de caractéristiques (Feature combination) : les données issues des deux techniques sont combinées et transmises à un seul algorithme de recommandation.
- Cascade : Cette méthode hybride se fait selon deux techniques : une première technique permet de générer un ensemble de candidats potentiels, et une deuxième technique permet de raffiner les recommandations. Cette méthode a pour avantage que si la première technique génère peu de recommandations, ou si ces recommandations sont ordonnées afin de permettre une sélection rapide, la deuxième technique ne sera plus utilisée [Naak, 2009].
- Par augmentation de propriétés (Feature augmentation) : le résultat d'une technique est utilisé comme entrée de l'autre technique [Burke, 2002].
- Méta-niveau (Meta-level) : Comme avec la méthode d'augmentation de caractéristiques, une première technique est utilisée, mais cette fois ci, non pas pour produire de nouvelles caractéristiques, mais pour produire un modèle. Et dans la deuxième étape, c'est tout le modèle qui servira d'entrée pour la deuxième technique [Burke, 2002].

4. Recommandation des papiers de recherche

4.1. Introduction

La recommandation est devenue de plus en plus importante et a changé la façon de communiquer entre les utilisateurs et les sites Web. Les systèmes de recommandation ont de nombreuses applications dans de nombreux domaines tels que l'économie, l'éducation, la recherche scientifique, etc.

Les systèmes de recommandation ont été introduits dans des communautés scientifiques pour récupérer de l'information de manière efficace [Amy *et al*.2013], [liu *et al*, 2015], [cao *et al*, 2017]. Dans la recherche scientifique, les systèmes de recommandation peuvent fournir des documents pour les chercheurs et les aider à trouver rapidement les documents dont ils ont besoin. Par exemple, pour les jeunes chercheurs ayant une expérience de publication limitée, les SR peuvent leur recommander des nouveaux articles ou des articles classiques relatifs à des domaines connexes pour élargir leurs horizons et intérêts de recherche. Au contraire, pour les chercheurs séniors avec une expérience de publication plus solide, les SRs recommandent principalement des papiers qui s'alignent sur leurs intérêts de recherche [Sugiyama *et al*, 2010] .

Recommander des articles scientifiques similaires pour les chercheurs est appelée dans la communauté scientifique recommandation scientifique ou recommandation des papiers de recherche.

Les SR des papiers de recherche visent à aider les chercheurs atténuer la surcharge d'information et trouver les documents pertinents en classant les publications et recommandant les articles top N associés aux intérêts ou objectifs de recherche d'un chercheur [Bai *et al*, 2018] .De nos jours, les systèmes de recommandation des papiers scientifiques sont devenus des outils indispensables dans le domaine académique. Ses algorithmes de recommandation sont continuellement mis à jour. Ainsi, l'exactitude de la recommandation s'améliore avec le temps. En plus, les systèmes de recommandation sont plus personnalisés et plus efficaces en les comparant à la technique traditionnelle de recherche par mots clés, en particulier dans le cas de quantités massives de données [Sun *et al*, 2014] , [liu *et al*,2015], [Sharma *et al*, 2017]. Le résultat de la recherche par mots-clés n'est pas toujours approprié, et le nombre d'items est relativement élevé [Bai *et al*, 2018]. C'est au chercheur de filtrer les résultats de la recherche pour obtenir les items dont il a besoin. Dans le cas de chercheurs différents, s'ils entrent la même requête, ils peuvent obtenir les mêmes résultats de recherche, parce que la technique de recherche par mots clés ne tient pas compte des différents intérêts et objectifs des utilisateurs. En outre, certains chercheurs ne savent pas formuler leurs besoins, ce qui entraîne la saisie de mots clés inappropriés. En comparaison, les systèmes de recommandation des papiers de recherche, généralement tiennent compte des intérêts des chercheurs, la relation de co-auteur et la relation de citation pour concevoir les algorithmes de recommandation et fournir la liste de recommandation. Il convient de noter que le nombre de résultats peut être court et contrôlable pour s'assurer que le SR soit personnalisé et efficace.

Depuis l'introduction des systèmes de recommandation, de nombreux algorithmes de recommandation sont apparus [Bai *et al*, 2018].Dans le domaine de recommandation des papiers de recherche, les techniques de recommandation utilisées sont divisées en quatre catégories principales : filtrage basé sur le contenu, filtrage collaboratif, les méthodes basées sur les graphes (GB) ainsi que les méthodes de recommandation hybrides comme illustré par la figure 5.

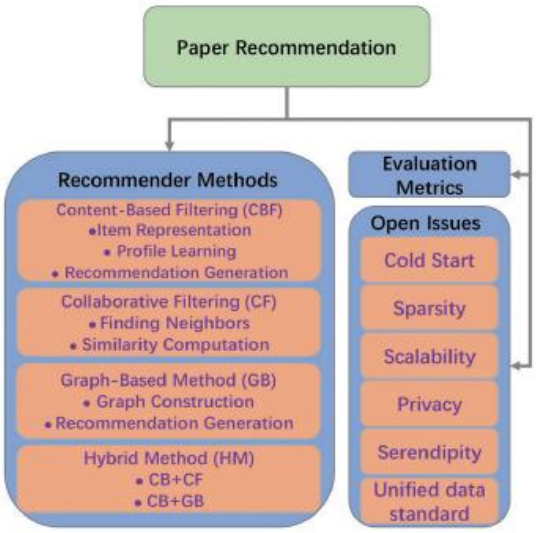

Figure 5: Recommandation des papiers **[**Bai *et al*, 2018]

4.2. Filtrage basé sur le contenu

Dans les systèmes de recommandation des papiers de recherche, les éléments sont les papiers dans la bibliothèque numérique et les utilisateurs sont les chercheurs. Dans la méthode CBF, d'abord les papiers du chercheur sont collectés. Les papiers d'un chercheur ou d'autres informations sont utilisés pour construire son profil.

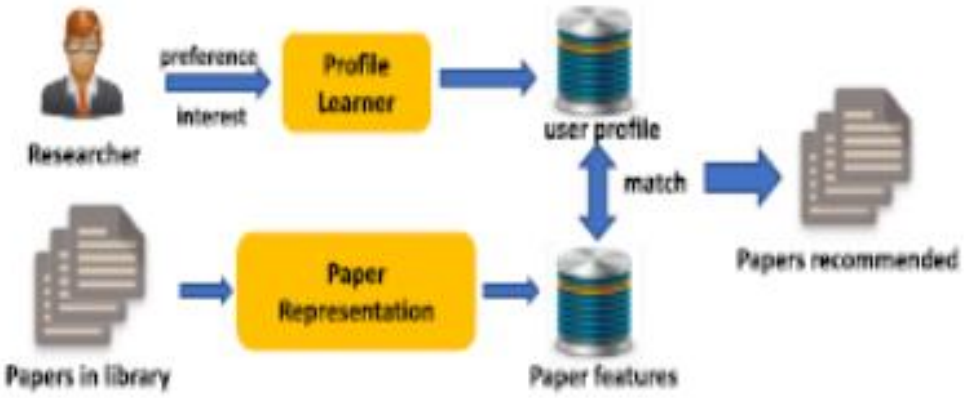

Figure 6: Système de recommandation des papiers basé sur le contenu [Bai *et al*, 2018]

Il existe de nombreuses façons pour construire le profil d'un chercheur [Bai *et al*, 2018]. Par exemple, les préférences du chercheur et ses intérêts peuvent être représentés par l'extraction de mots-clés du domaine de recherche du chercheur. De plus, les systèmes de recommandation des papiers de recherches peuvent extraire des mots-clés du titre, du résumé et du contenu des papiers pour représenter ces papiers. Les papiers candidats peuvent être récupérés à partir de la bibliothèque numérique. Le SR calcule ensuite la similarité des mots clés entre le profil du

chercheur et les papiers candidats puis les trie. Enfin, les papiers candidats ayant une grande similarité seront recommandés au chercheur.

Selon la logique sous-jacente. Le système CBF extrait l'information sur les papiers et les compare. Si le papier est lié aux intérêts du chercheur, il sera découvert. En outre, par rapport aux moteurs de recherche basés sur des mots clés, CBF considère généralement les intérêts actuels du chercheur, et n'implique pas d'autres chercheurs. Si les intérêts des chercheurs changent, les résultats changeront également. La figure 6 montre la structure générale des systèmes de recommandation fondés sur le contenu.

De la figure 6, nous pouvons voir la progression de la recommandation du CBF comprenant trois étapes principales : Représentation des items, apprentissage du profil et génération de recommandations.

4.2.1 Représentation des items

Dans la pratique, les items ont généralement besoin de certains attributs spéciaux pour se distinguer. Ces attributs peuvent être divisés en deux catégories principales : attributs structurés et attributs non structurés [Bai *et al*, 2018]. Pour les attributs structurés, la valeur de l'attribut est limité et spécifique. Par contre pour les attributs non structurés, la valeur d'attribut est souvent moins claire. Parce que sa valeur est illimitée, et ne peut pas être directement utilisée pour analyse.

Dans le domaine de recommandation des papiers de recherche, les structures globales des papiers sont similaires, mais leurs contenus sont illimités, et chaque auteur a son propre style d'écriture. Afin de représenter tous les papiers et de calculer la similarité entre eux, nous devons traduire le contenu des documents en éléments structurés. Depuis, les systèmes de recommandation de papier ont été proposés, il y a beaucoup de méthodes de représentation des items, comme le modèle TF-IDF [Jomsri *et al*, 2010], le modèle d'extraction de phrases clés, le modèle de langage et ainsi de suite [Bai *et al*, 2018].

Le modèle TF-IDF (de l'anglais term frequency-inverse document frequency) a été fréquemment utilisé pour la recherche d'informations et l'extraction de textes [Jomsri *et al* ,2010]. La valeur de TF-IDF est une mesure statistique pour évaluer l'importance d'un mot pour un document dans une collection ou un corpus. L'idée de base du modèle de TF-IDF est divisée en deux aspects. D'un côté, plus le mot clé K apparaît dans le document D, le plus K est important pour le document D. D'autre part, plus K apparaît avec une grande fréquence dans différents documents, moins K est important pour distinguer les documents. L'équation (1) est définie comme suit [Sugiyama *et al*, 2010] :

$$w_{t_k}^{p_{rec}} = \frac{tf(t_k, P_{rec})}{\sum_{s=0}^{m} tf(t_s, P_{rec})} \times \log \frac{N}{df(t_k)} \quad (1)$$

Où tf (t_k, P_{rec}): fréquence du mot-clé t_k dans le papier p

N : Nombre de papiers dans l'ensemble candidat

df(t_k):) fréquence d'occurrence du mot-clé t_k

CBF utilise le modèle TF-IDF pour calculer les vecteurs de caractéristiques f^{Prec} de chaque papier candidat [Sugiyama *et al*, 2010], [Bai *et al*, 2018]. Ces vecteurs peuvent déterminer la pertinence d'un papier de recherche pour le chercheur
La définition de f^{Prec} est donné par l'équation (2) la suivante :

$$f^{Prec} = (w_{t1}^{Prec}, w_{t2}^{Prec}, \dots\dots, w_{tm}^{Prec}) \quad (2)$$

où m est le nombre de termes distincts dans le document, et t_k(k=1, 2 ,…,m) désigne chaque terme, deux vecteurs pour chaque papier sont utilisés comme différentes requêtes d'entrée. Ce modèle est populaire pour les systèmes de recommandation CBF, de nombreux chercheurs ont adopté une version modifiée dans leur recherche. Certains chercheurs réalisent que lorsque nous lisons un papier scientifique, nous pouvons être curieux sur le problème traité dans le papier ou la solution apportée au problème. Ainsi, ils utilisent le modèle TF-IDF, le modèle thématique et le modèle thématique basé sur les concepts pour calculer la similarité et trouver pour les utilisateurs les papiers les plus axés sur les problématiques et les papiers les plus axés sur les solutions, qui répondent séparément à l'objectif spécifique du chercheur en matière de lecture [Bai *et al*, 2018].

Outre le modèle TF-IDF, un modèle d'extraction de phrases-clés (généralement constitué d'un à trois mots) est utilisé pour produire une riche description du contenu des papiers [Bai *et al*, 2018]. La liste des phrases-clés est une courte liste de mots clés qui reflètent le contenu d'un papier, capture les principaux thèmes discutés et fournit un bref résumé de son contenu. Dans ce modèle, le titre, le résumé et les mots-clés d'un article sont représentés par différents vecteurs : $V_{résumé}$,V_{titre},$V_{mots\text{-}clés}$ respectivement [Basu *et al*, 2012].

Le vecteur de mots-clés est extrait de la section « mot-clés » du papier. Si le papier ne comporte pas la section « Mot-clés », le système d'analyse considérera les mots les plus représentatifs comme les mots-clés nécessaires [Bai *et al*, 2018].

4.2.2 Apprentissage du profil

Les systèmes de recommandation CBF supposent que le chercheur a noté « J'aime » ou « J'aime pas » sur certains items et a publié des papiers selon ses intérêts individuels. L'objectif de cette étape est de générer le modèle du profil selon les actions historiques des chercheurs. Étant donné que le profil du chercheur inclut la direction de recherche des chercheurs, les systèmes peuvent déterminer si le chercheur U aime un nouvel item par ce modèle [Chen *et al*, 2007].

Il est évident que le profil du chercheur doit s'appuyer sur l'information générée par le chercheur. Diverses méthodes existent pour construire des profils utilisateurs. Une des méthodes consiste à construire le profil utilisateur avec un mélange de sujets extraits des publications antérieures du chercheur par l'algorithme LDA (Allocation de Dirichlet latente)[Bai *et al*, 2018]. Les vecteurs $V_{résumé}$, V_{titre}, $V_{\textit{mots clés}}$, sont extraits des papiers des actions historiques du chercheur pour construire le profil. Le profil de l'utilisateur pourrait être mis à jour si le chercheur publie ou évalue de nouveaux papiers dans le futur.

Le système d'information basé annotations (tag-based) utilise un analyseur de préférences de l'utilisateur nommé *« User Preference Crawler »* pour parcourir les données des préférences de l'utilisateur. Le profil de l'utilisateur est construit par les papiers postés par chaque utilisateur individuel et un ensemble d'annotations postées par les utilisateurs [Jomsri *et al*, 2010], [Bai *et al*, 2018]. De même, les annotations et l'ensemble des documents annotés par les chercheurs peuvent être exploités par le module d'extraction de phrases- clés pour construire le profil de l'utilisateur [Bai *et al*, 2018].

Pour faciliter la personnalisation des systèmes de recommandation, les jeunes chercheurs qui ont publié quelques articles et les chercheurs seniors avec de nombreuses publications pourraient être différenciés [Sugiyama *et al*, 2010], [Bai et al , 2018].

Pour un papier, le vecteur caractéristique f^{Prec} est initialement défini par le TF dans le modèle TF-IDF. La définition de f^{Prec} est identique à l'équation (2).

Après avoir obtenu les vecteurs caractéristiques des papiers, la construction du profil de l'utilisateur est divisée en deux catégories : les jeunes chercheurs et les chercheurs séniors. Pour les jeunes chercheurs n'ayant qu'un seul papier *p1*, la construction du profil de l'utilisateur P_{user} ajoutera la contribution des articles cités par *p1*. Pour les chercheurs séniors avec plusieurs articles publiés dans le passé $p_i (i = 1, \ldots, n-1)$, le profil de l'utilisateur ajoutera la contribution des papiers citant pi et les papiers dans la liste de références de p_i. Cette méthode rend les profils des jeunes chercheurs et même séniors plus spécifiques.

Toutes ces méthodes introduites de l'apprentissage des profils dépendent de l'enregistrement de l'historique du chercheur ou ses actions. Dans certains systèmes de recommandations, pour

déterminer le profil de l'utilisateur, ils tiennent compte des papiers fournis par le chercheur comme entrée pour construire le profil de l'utilisateur [Bai *et al*, 2018]. Après la fourniture du papier, les informations nécessaires pour le système seront extraites des parties du papier : titre de l'article, l'introduction, les travaux en relation, la conclusion, les références. En outre, pour satisfaire l'objectif de lecture spécifique de l'utilisateur, le résumé est parfois divisé en deux parties : la description du problème et la description de la solution afin que le système puisse recommander des articles sur deux aspects respectivement [Bai *et al*, 2018].

De plus, il existe d'autres formulaires pour représenter le profil de l'utilisateur *Docear* est un système de recommandation qui possède la caractéristique unique de l'utilisation des cartes mentales pour la gestion de l'information [Beel e*t al* , 2011]. Les utilisateurs de *Docear* organisent leurs données dans une structure de données arbre, et ils construisent un modèle utilisateur à partir de la collection de cartes mentales de l'utilisateur pour correspondre à sa bibliothèque numérique.

4.2.3 Génération de recommandations

Les représentations des papiers candidats et les profils des chercheurs sont conçus pour sélectionner les N éléments les plus pertinents aux utilisateurs. La pertinence des attributs des chercheurs par rapport aux attributs des articles peut être obtenue par la mesure de similarité telle que la similarité cosinus.

Etant donné deux vecteurs d'attributs A et B, la similarité cosinus est calculée comme suit [Jomsri et all ,2010]:

$$\boldsymbol{Similarity} = \cos(\boldsymbol{\theta}) = \frac{\boldsymbol{A} \,.\, \boldsymbol{B}}{||\boldsymbol{A}||.||\boldsymbol{B}||} \qquad \textbf{(3)}$$

La recommandation des papiers utilise des vecteurs des profils des utilisateurs P_{user} et le vecteurs de caractéristiques des papiers candidats F^{Prec} , qui sont définis avant de calculer la similarité cosinus de P_{user} et F^{Prec} en utilisant l'équation (3) [Sugiyama *et al*, 2010].

Certaines recherches antérieures fournissent non seulement aux chercheurs les papiers les plus pertinents, mais aussi permettent de fournir des recommandations aléatoires avec les papiers de domaines lointains et variés. La recommandation hasardeuses est utile pour les chercheurs pour découvrir de nouvelles idées, approches ou façons de penser [Bai *et al*, 2018].

Après le calcul de similarité entre le profil de l'utilisateur et les papiers candidats, une liste de résultats sera générée. La dernière étape des systèmes de recommandation et le classement des résultats dans un certain ordre. Ainsi, la liste finale des N meilleures papiers sera recommandée au chercheur. En classant les articles candidats, le nombre d'articles les citant est parfois

considéré comme critère. Par la suite, les chercheurs peuvent utiliser ce système de recommandation pour trouver des papiers qui les intéressent.

Cependant, il y a certains problèmes dans les systèmes de recommandation de type CBF. D'un côté, le CBF ne prend pas en considération la qualité telle que la source, le style parce que ses techniques se basent uniquement sur l'analyse de mots. D'autre part, il y a le problème du nouvel utilisateur. Si un jeune chercheur sans beaucoup d'expérience de recherche utilise le système, ce dernier peut fonctionnera de façon inefficace. Parce qu'il ne peut pas extraire suffisamment d'informations de l'utilisateur et de ce fait la liste recommandée peut ne pas être fiable [Bai *et al*, 2018].

4.3. Filtrage collaboratif (CF)

Comme les techniques de recommandation de CBF, CF a besoin de connaître les intérêts des utilisateurs, ce qui est particulièrement efficace pour recommander des articles connexes. L'idée de base de CF est que si les utilisateurs A et B évaluent certains éléments communs, leurs intérêts seront considérés comme similaires. S'il existe des éléments dans l'enregistrement de l'utilisateur B mais pas dans celui de l'utilisateur A, ces éléments peuvent être recommandés à l'utilisateur A. En d'autres termes, CF est le processus de recommandation d'éléments en utilisant les opinions d'autres utilisateurs. Les notes ou opinions peuvent être obtenues à partir de certains sites Web de gestion des références sociales comme CiteULike, ou en demandant aux utilisateurs de remplir un questionnaire [Bai *et al*, 2018].

Le système de filtrage collaboratif localise l'utilisateur pair en tenant compte de son historique de notation et en trouvant l'utilisateur similaire. Ensuite, CF utilise le voisinage pour générer la recommandation. Les systèmes de recommandation des FC ont généralement besoin d'une matrice item-utilisateur pour représenter les notes ou les commentaires des utilisateurs sur les items. Les évaluations peuvent être utilisées pour représenter les intérêts des utilisateurs. Après avoir construit la matrice, le système calculera la similarité entre les utilisateurs pour trouver des utilisateurs similaires appelés «utilisateurs voisins» afin de recommander des articles. Une matrice item-utilisateur est présentée dans le tableau 1, les éléments de la matrice sont les évaluations des utilisateurs. Dans cette matrice, les notes sont 0 et 1, les notes peuvent prendre plus de valeurs pour exprimer les différents degrés d'appréciation ou d'aversion. La structure générale des systèmes de filtrage collaboratifs est illustrée par la figure 7.

	Item1	Item2	Item3	Item4	…….	ItemX
Utilisateur1	0	1	0	1	…	0
Utilisateur2	1	1	0	0	…	0
Utilisateur3	1	0	1	1	…	0
……	….	…	…	…	…	…
UtilisateurY	1	0	1	1	…	1

Tableau 1: Matrice Utilisateur-Item**.**

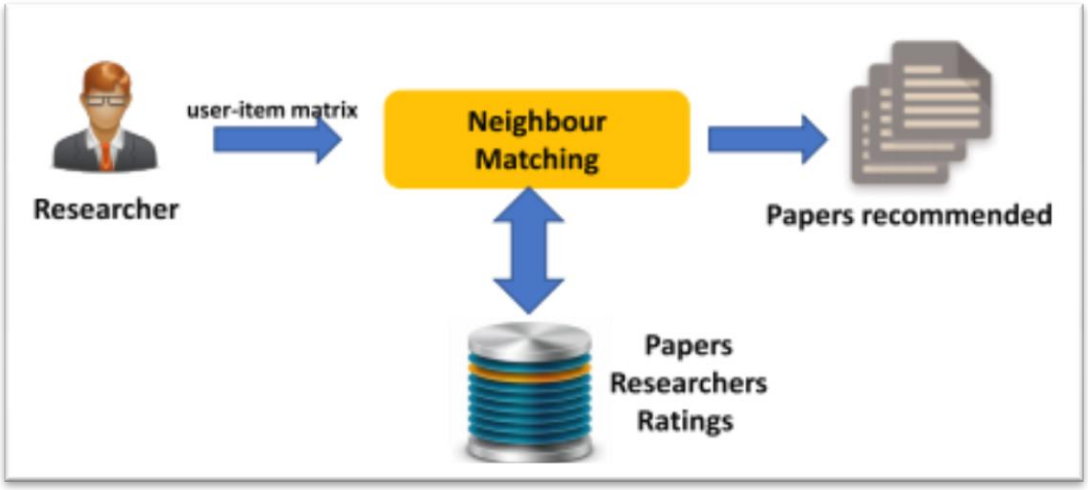

Figure 7:système de filtrage collaboratif pour la recommandation des papiers.

(d'après [Bai *et a*l,2018])

Par rapport à la méthode de filtrage basé sur le contenu, CF présente des avantages différents: le contenu de l'article recommandé n'est pas pris en compte, car la méthode de recommandation dépend des évaluations faites par les utilisateurs et ne tient pas compte des types d'articles auxquels ils appartiennent. En outre, les articles recommandés aux utilisateurs peuvent ne pas être pertinents pour la recherche actuelle de l'utilisateur, car la similarité est mesurée entre les relations entre les utilisateurs.

CF contient principalement les deux catégories de méthodes : les approches basées sur l'utilisateur et les approches basées sur les items. Selon les besoins différents des utilisateurs, ces techniques de recommandation peuvent collecter les données nécessaires et recommander des papiers. Les métadonnées de CiteULike peuvent être utilisées pour exécuter l'algorithme de

recommandation CF, les métadonnées contiennent de nombreux utilisateurs et leurs annotations uniques sur les papiers [Bai *et al* , 2018].

L'algorithme de recommandation est classique et simple: dans le filtrage basé sur l'utilisateur, l'utilisateur cible est mis en correspondance avec les données collectées pour trouver les voisins qui ont des enregistrements similaires. Une fois les voisins trouvés, tous les papiers ayant une préférence dans l'historique du voisin seront considérés comme des papiers candidats à recommander aux utilisateurs cibles. Dans le filtrage basé sur les items, le système recommande les papiers en faisant correspondre les papiers avec les enregistrements historiques de l'utilisateur cible. Pour le CF basé sur l'utilisateur, la similarité entre deux utilisateurs est calculée par les notes de leurs articles communs [Parra et Brusilovsky, 2010]. L'équation est la suivante:

$$\boldsymbol{Sim(u,n)} = \frac{\sum_{i \subset \boldsymbol{R_{u,n}}} (\boldsymbol{r_{ui}} - \bar{\boldsymbol{r}}_{\boldsymbol{u}})(\boldsymbol{r_{ni}} - \bar{\boldsymbol{r}}_{\boldsymbol{n}})}{\sqrt{\sum_{i \subset \boldsymbol{CR_{u,n}}} (\boldsymbol{r_{ui}} - \bar{\boldsymbol{r}}_{\boldsymbol{u}})^2} \sqrt{\sum_{i \subset \boldsymbol{CR_{u,n}}} (\boldsymbol{r_{ni}} - \bar{\boldsymbol{r}}_{\boldsymbol{u}})^2}} \qquad (4)$$

où r sont les évaluations, u est l'utilisateur cible et n est l'utilisateur voisin, r_{ui} représente les évaluations données par l'utilisateur u à l'item i, $\bar{r}$ est la note moyenne de l'utilisateur u sur tous ses items. $CR_{u,n}$ représente l'ensemble commun d'items entre l'utilisateur u et l'utilisateur n. Les papiers des utilisateurs voisins sont recommandés à l'utilisateur cible en classant les notes prévues pour l'utilisateur cible u. Les relations sociales sont généralement ajoutées pour trouver les bons voisins. Après avoir trouvé les voisins les plus proches, l'étape suivante consiste à prédire la note de l'utilisateur cible u pour l'article i. La formule est la suivante [Bai *et al*, 2018]:

$$\mathbf{pred(u,i)} = \overline{\mathbf{r_u}} + \frac{\sum_{\mathbf{n} \subset \mathbf{neigh(u)}} \mathbf{userSim(u,n)}.(\mathbf{r_{ni}} - \overline{\mathbf{r_n}})}{\sum_{\mathbf{n} \subset \mathbf{neigh(u)}} \mathbf{userSim(u,n)}} \qquad (5)$$

Pour une matrice d'item-utilisateur donnée, le modèle de factorisation matricielle joue un rôle important dans les systèmes de recommandation de filtrage collaboratif. Le modèle de factorisation matricielle est utilisé pour prédire les notes des articles candidats. Les algorithmes CF basés sur utilisateur recommandent des articles dans le système d'annotations sociales [Parra et Brusilovsky, 2010]. Les chercheurs résument le processus de filtrage collaboratif basé utilisateur en deux étapes : la première étape consiste à trouver les voisins de l'utilisateur cible, la deuxième étape consiste à utiliser les voisins pour classer les éléments, puis à recommander les N premiers papiers à l'utilisateur [Parra et Brusilovsky, 2010]. Pour améliorer la qualité du résultat recommandé, les deux étapes sont améliorées.

De plus, les articles scientifiques sont recommandés en utilisant les relations sociales telles que les amis, la popularité de la recherche .En outre, le profil de l'utilisateur, le profil du groupe et les

relations sociales entre les utilisateurs sont généralement considérés pour recommander les papiers scientifiques.

Semblable à l'approche basée sur l'utilisateur, le filtrage collaboratif basé sur les items comprend les deux étapes: le calcul de similarité et la génération de prédiction [Sarwar *et al*, 2001] .Dans un premier temps, la similarité telle que la similarité cosinus ou la similitude thématique des articles cibles i avec l'ensemble des articles évalués par l'utilisateur cible sont utilisés pour trouver les k articles les plus similaires i_1, i_2,. . . , i_k pour l'ensemble d'articles candidats. Dans la deuxième étape, après avoir obtenu les papiers les plus similaires, la prédiction serait ensuite calculée par une moyenne pondérée des notes de l'utilisateur cible sur ces éléments similaires. Pour garantir la pertinence du résultat, un système amélioré de filtrage collaboratif basé sur les items recommande des articles évalués par les connexions de l'utilisateur cible U. Ainsi, Les articles recommandés sont non seulement similaires à la publication cible P d'intérêt pour l'utilisateur cible U, mais sont également populaire parmi les connexions de l'utilisateur cible U. Dans ce système, les chercheurs trouvent d'abord les connexions de l'utilisateur cible qui échangent et partagent des références bibliographiques avec lui. Ensuite, des facteurs de corrélation de mots sont utilisés pour déterminer les articles candidats qui sont similaires à l'article cible P. Finalement, le système recommande les papiers avec les scores les plus élevés à l'utilisateur cible U [Bai *et al*, 2018].

4.4. Méthode basée sur les graphes (GB)

Comme son nom l'indique, la méthode fondée sur les graphes met principalement l'accent sur la construction de graphes. Le graphe peut être construit par les réseaux de citation, les réseaux sociaux, ... etc. Les chercheurs et les papiers sont les différents nœuds du graphe.

Les relations entre chercheurs (i.e. auteurs), chercheurs et papiers, papiers et papiers peuvent être considérés comme des arcs entre les nœuds. Ensuite, le système de recommandation peut utiliser un algorithme comme celui de la marche aléatoire (Random Walk algorithm) sur les graphes pour trouver les papiers pertinents pour les chercheurs. L'avantage of GB est que cette méthode peut utiliser des informations de différentes sources pour la recommandation.

Les techniques CBF et CF utilisent un ou deux types d'informations. La méthode GB peut ajouter des relations sociales ou des relations entre chercheurs dans le SR pour améliorer le résultat de la recommandation.

L'utilisation de la structure d'un graphe pour recommander les papiers est une nouvelle méthode. Le GB principalement utilise les relations entre les nœuds.

Dans le modèle basé graphe, initialement, nous avons besoin de collecter les données concernant les chercheurs et les papiers. Puis, le système les représente avec un graphe hétérogène $G(V; E)$, où $V = V_U \cup V_P$, V_U représente les chercheurs dans le système et V_P est l'ensemble des papiers publiés ou référencés par les chercheurs.

Pour chaque tuple $(U; P)$, Il existe un arc $E(v_u, v_p)$ dans le graphe, et $v_u \in V_U$, $v_p \in V_P$.

En plus, dans quelques systèmes de recommandation à base de graphes, il existe aussi des arcs tels que $E(v_u, v_u)$, $E(vp, vp)$ qui signifient la considération des relations entre chercheurs et aussi les relations entre papiers.

Dans le modèle à base de graphes, l'activité de recommandation des papiers se transforme en une tâche de recherche sur les graphes [Huang *et al*, 2002].

La recommandation dans un SR à base de graphes peut être résumée en deux étapes : la construction du graphe et la génération de la recommandation.

4.4.1 Construction du graphe

Pour la recherche académique, les chercheurs lisent et recherche des papiers pertinents de quelques bibliothèques numériques telles que IEEE Xplore et CiteULike. Les chercheurs peuvent Collecter les données concernant les utilisateurs et les papiers à partir des sites web mentionnés ci-dessus pour construire le graphe.

Par exemple, la relation entre un chercheur et un papier signifie que le chercheur est intéressé par ce papier.

Une matrice W_{RA}^{nxm} (i,j) est utilisé pour indiquer si un chercheur R_i est intéressé par un article A_j comme indiqueé par l'équation (6).

R est l'ensemble de n chercheurs $R_1, R_2, \ldots\ldots R_n$. A est l'ensemble de m articles $A_1, \ldots\ldots, A_m$.

La relation d'auteurs commun (co-auteur) est également ajouté dans le graphe basique [liu *et al*, 2015], [Bai et all,2018]. Pour les relations de co-auteur entre les papiers, une autre matrice W_{AA}^{mxm} (i,j) est utilisée pour indiquer si deux articles A_i et A_j ont un ou des auteur(s) communs comme illustré par l'équation (7).

$$W_{RA}(i,j) = \begin{cases} 1 & \text{si } R_i \text{ intéressé par } Aj \\ 0 & \text{sinon} \end{cases} \quad (6)$$

$$W_{AA}(i,j) = \begin{cases} 1 & \text{si Ai et Aj ont un auteur commun} \\ 0 & \text{sinon} \end{cases} \quad (7)$$

Après avoir construit les deux matrices mentionnées, elles seront transformées en graphes.

Soit G = ($VR \cup VA$, $ERA \cup EAA$), avec $ERA \subseteq VR \times VA$, et $EAA \subseteq VA \times VA$.

- *VR* et *VA* sont les ensembles des nœuds représentant les chercheurs et les papiers

- *EAA* représente l'ensemble des relations d'intérêt entre les chercheurs et les papiers.
- *ERA* représente les relations de type co-auteur.

Si $W_{RA}(i; j)$ vaut 1,cela signifie qu'il existe dans le graphe un arc entre le chercheur « i » *et le papier « j »* De manière similaire, si $W_{AA}(i, j)$ vaut 1, donc il existe un arc entre le papier i et le papier j .

Un graphe hybride avec les relations co-auteur peut être construit, qui sera utilisé pour générer des recommandations. Un autre graphe hétérogène appelé graphe bi-relationnel (bRG) peut être utilisé pour recommander des papiers [Tian et Jing, 2013]. bRG est similaire aux graphes précédemment mentionnés, il inclut aussi les chercheurs et les papiers. En plus, bRG contient un sous graphe de similarité de papiers, un sous graphe de similarité de chercheurs et un graphe biparti connectant les chercheurs et les papiers. Il y a également un autre type de graphe : graphe de citations ou réseau de citations.

Le graphe de citations contient les papiers et la relation de citation entre papiers. Les nœuds représentent les différents papiers dans le réseau de citations, et les arcs représentent les relations de citation entre papiers. L'idée de base dans le graphe de citation est que si deux papiers ont des références communes ou bien ils sont cités par un papier, ils sont considérés similaires [Bai *et al*, 2018]. Par conséquent, la recommandation peut être donnée en analysant la structure du réseau de citations. Dans le SR proposé par [Steinert et al, 2015], tous les papiers considérés comme $D = p_1, p_2, \ldots, p_n$ pour construire le graphe de citations. *Rp* est un sous ensemble de D, *Rp* contient tous les papiers cités par le papier p .Ainsi, les papiers dans *Rp* sont reliés au papier p. si un papier p_k dans *D* est relié à un ou plusieurs papiers dans Rp, alors le papier p_k sera recommandé à l'utilisateur. En se basant sur des idées similaires, une méthode est proposée pour recommander des papiers en utilisant le réseau des citations et un algorithme à base de contenu. Dans un graphe hétérogène pondéré, les chercheurs remplacent la partie auteur avec un graphe de mots clés contenant les mots-clés extraits de chaque papier en utilisant le modèle TF-IDF. Le poids de la relation de citation entre une paire de papiers c'est la similarité cosinus entre deux vecteurs pi et pj. Le score TF-IDF c'est le poids du mot-clé au papier, et la similarité de deux termes c'est le poids des arcs.

De plus, les relations de type co-auteurs entre les auteurs peuvent être ajoutées dans le réseau de citations. Ce graphe est appelé réseau collaborative des citations. Il possède les trois différents types de liens représentant différentes relations : relations de citation, relations collaboratives et les relations auteur-papier [Wang al, 2016].

La principale forme de construction de graphe a été introduite au-dessus. Il existe d'autres types de graphes utilisés pour générer des papiers pertinents pour les chercheurs ou un papier donné parmi un ensemble de papiers candidats, comme les cartes conceptuelles. [Ohta *et al*, 2011]

4.4.2 Génération de la recommandation

Les algorithmes dans les systèmes de recommandation à base de graphes généralement ne considèrent pas les caractéristiques du contenu du papier ni le profil du chercheur. La raison est qu'ils ne conviennent pas pour être représentés comme nœuds de graphe pour la recommandation. Dans le graphe les chercheurs et les papiers représentent les deux types de nœuds. Le système de recommandation profite des informations de la structure du graphe pour trouver les papiers pertinents. L'algorithme de marche aléatoire avec redémarrage peut être utilisé pour classer les articles [liu *et al*, 2015], [Tian et Jing, 2013]. Le principe de l'algorithme classique de marche aléatoire est qu'un marcheur aléatoire est utilisé pour traverser un graphe à partir d'un ou d'une série de nœuds avec une probabilité « a » d'aller aux nœuds voisins du nœud courant et une probabilité « 1-a » de sauter aléatoirement à n'importe quel nœud dans le graphe. Chaque marche donne une distribution de probabilité qui indique la probabilité que chaque nœud dans le graphe est accessible. Cette distribution de probabilité est utilisée comme entrée pour la prochaine marche et ce processus se répète itérativement. Quand certaines pré-conditions soient satisfaites, la distribution tend vers la convergence. La méthode de marche aléatoire avec redémarrage est une amélioration de l'algorithme basique marche aléatoire. Quand le marcheur commence à partir d'un nœud dans le graphe, il a une probabilité « a » d'aller aux nœuds voisins à partir du nœud courant et une probabilité « 1-a » de retourner sur le nœud source [Fouus *et al*, 2007] . Le graphe bi-parti utilise l'algorithme de marche aléatoire avec redémarrage pour calculer le classement des papiers [Bai *et al*, 2018].

En plus, l'algorithme de marche aléatoire est utilisé parfois dans les domaines connexes des systèmes de recommandation. Par exemple, son utilisation pour trouver les utilisateurs similaires pour un utilisateur cible [Xu *et al*, 2016]. Dans l'étude, initialement les chercheurs utilisent les relations sociales pour construire un réseau entre utilisateurs. Pour les utilisateurs cibles, l'hypothèse est qu'ils ont tendance à accepter la recommandation de leurs amis avec des intérêts similaires. Le modèle de marche aléatoire est utilisé pour obtenir des utilisateurs similaires. Ensuite, les systèmes prédisent les notes par les utilisateurs les plus similaires. Enfin, la liste des recommandations est générée.

L'algorithme « PaperRank » est largement utilisé dans les systèmes de recommandation pour calculer la pertinence entre les articles dans le réseau de citations [Bai *et al*, 2018]. PaperRank

est l'extension du modèle PageRank pour évaluer les articles scientifique, en considérant les relations indirectes entre papiers. L'analyse de citation dans les méthodes précédentes est simple : ISI Journal Impact Factor calcule une moyenne des fréquences de citations des articles publiés et retourne une liste triée des journaux [Garfield, 1972] . En fait, le nombre des papiers cités est utilisé pour classer les papiers selon le nombre de relations de citations directes. Le principe de l'algorithme PaperRank est l'utilisation des papiers pour remplacer les pages dans PageRank [Haveliwala, 2003].Chaque valeur individuelle de PageRank peut être calculé par l'équation suivante :

$$\mathbf{PR(P_i) = \frac{1-d}{N} + d \sum_{i \neq j} \frac{PR(P_i).l(P_i,P_j)}{L(P_j)}} \qquad (8)$$

Où $P_1, P_2, \ldots\ldots, P_N$ sont les N papiers dans le réseau de citations,

$PR(P_i)$ c'est la valeur PageRank du papier P_i (le score de classement du papier),

$L(P_i)$ c'est le nombre de références du papier P_i .

d est le coefficient d'amortissement,

$l(P_i; P_j)$ est la fonction qui représente si l'article P_i a cité l'article P_j.

si P_j est cité par P_i alors l(Pi, Pj) vaut 1, sinon l (Pi, Pj) vaut 0.

En utilisant cette méthode, l'importance des articles individuels peut être exprimée.

4.5. Les méthodes hybrides (HM)

Pour améliorer l'exactitude des résultats de recommandation et obtenir une meilleure performance, quelques système de recommandation des papiers scientifiques combinent deux ou plusieurs technique de recommandation afin de recommander des papiers personnalisés aux chercheurs [Bai *et al*, 2018].

Evidemment, l'avantage de HM est la possibilité d'utiliser la combinaison de différentes techniques de recommandations ainsi que différentes sources d'informations. Dans la suite de cette section, nous présentons quelques techniques de recommandation hybrides.

4.5.1. Filtrage collaborative + basé sur le contenu:

Les deux méthodes de recommandation à base de contenu ainsi le filtrage collaboratif ont leur propres avantages et inconvénients .Quelques études essayent de combiner les deux méthodes avec différentes formes pour rendre meilleure la recommandation des papiers et surmonter leurs lacunes telles que le problème de démarrage à froid et celui de la parcimonie [Sun *et al* ,2014], [Winoto *et al*,2012], [Sugiyama et Kan, 2013].

Les techniques à base de contenu construisent le profil du chercheur en capturant leurs intérêts de recherche antérieurs incarnés dans leurs publications passées. Les techniques de filtrage collaboratif visent à découvrir les documents de citation potentiels.

Le processus de recommandation d'articles comprend trois étapes. Initialement, les chercheurs doivent créer le profil de l'utilisateur à partir de ses articles publiés en utilisant le schéma **TF**.

Et calcule le vecteur des caractéristiques pour chaque papier candidat par le schéma TF-IDF . Ils trouvent *N* papiers avec les plus hauts scores de similarité cosinus. Deuxièment, pour les papiers, l'algorithme CF fonctionne sur la matrice de citation-papier en se basant sur l'idée que les articles similaires ont des citations similaires pour trouver les papiers potentiels. Le coefficient de corrélation de Pearson entre les vecteurs de citation au papier cible est utilisé pour mesurer la similarité. Les papiers avec la similarité la plus élevée avec le papier cible forment les papiers voisins. Finalement, la similarité cosinus du contenu sera calculée [Sun *et al* ,2014].

En combinant les deux méthodes, le système donne une performance supérieure aux systèmes de recommandation classiques. En se basant sur les techniques de recommandation traditionnelles, certains algorithmes modifiés sont apparus tels que l'algorithme CBF-Separé, l'algorithme CF-CBF séparés et l'algorithme CBF-CF parallèles. Tous ces algorithmes sont prouvés être meilleures par rapport à l'utilisation d'une technique unique de recommandation.

En plus, il existe que quelques méthodes hybrides spéciales comme le modèle de filtrage collaboratif avec facteur latent, le modèle de thèmes probabiliste. La performance de ces méthodes hybrides sont meilleures que les méthodes de base [Bai *et al*, 2018].

4.5.2. Basé sur le contenu+ basé sur un graphe :

La combinaison des méthodes à base de contenu et les méthodes basées graphes fonctionne mieux que les méthodes de recommandation classiques. Parce que la méthode basée sur le contenu peut obtenir le profil utilisateur à partir du contenu des articles qui intéressent les utilisateurs. La méthode à base de graphe peut utiliser le réseau de citations ou le graphe biparti pour trouver des papiers candidats plus potentiels à partir de la structure du graphe.

Les techniques à base de contenu avec le réseau de citations ont la capacité de recommander les articles les plus pertinents d'une bibliothèque numérique [Beel *et al*, 2017]. Le graphe biparti comprend les deux couches: la couche des papiers relie les papiers avec des relations de citation.

La couche des chercheurs relie les chercheurs à leurs relations sociales. En particulier, pour faire la recommandation avec plus de précision, une nouvelle méthode hybride de recommandation d'articles intégrant les informations sociales est proposée [Wang *et al*, 2016].

La méthode de recommandation comprend les trois types de relations:

1. Pour les chercheurs A et B, la confiance de base est que le chercheur A et le chercheur B se chevauchent dans leur bibliothèque.
2. La valeur du chercheur B sera augmentée si le chercheur B est l'auteur de quelques articles dans la bibliothèque de recherche de A.
3. est-ce que le chercheur A fait confiance aux connaissances du chercheur B dans un sujet spécial.

Les papiers candidats (CP) proviennent de la structure du graphe biparti. Le système de recommandation sélectionne CP dans les bibliothèques des chercheurs actuels. En construisant le profil des chercheurs, les jeunes chercheurs et les chercheurs séniors sont distingués. Les intérêts des jeunes chercheurs et les chercheurs séniors sont représentés par vecteurs des caractéristiques à travers le modèle TF-IDF pour analyser le contenu des papiers. Le classement du CP tiendra compte de la similarité entre les vecteurs caractéristiques des CP, le profil des chercheurs, la valeur de la confiance entre les propriétaires des CP, le chercheur actuel, le nombre de citations du CP et la réputation des auteurs.

En plus d'être combinés, les méthodes de recommandation peuvent être utilisées séparément. La méthode basée sur le contenu en utilisant le modèle TF-IDF obtient les vecteurs de caractéristiques des papiers candidats. La similarité est obtenue en calculant la similarité cosinus des articles candidats et l'article dans la cible de l'utilisateur. La méthode à base de graphes utilisant le réseau de citation classique exécute des algorithmes pour obtenir la préférence de l'utilisateur et recommander les N meilleurs papiers à l'utilisateur. L'approche hybride utilise les listes de résultats du deux méthodes mentionnées et leur donne un poids différent. Soit $f_{content}$ le résultat de la méthode basée sur le contenu, f_{graph} est le résultat de la méthode basée sur les graphes. Le résultat hybride f_{hybrid} est calculé comme suit

$$f_{hybrid} = w * f_{content} + (1 - w) * f_{graph} \qquad (9)$$

où w et (1- w) représentent les poids des deux méthodes.

La combinaison peut résoudre le problème de la sur-spécialisation et le problème des nouveaux items des méthodes classiques.Les méthodes hybrides ont de nombreuses combinaisons différentes et elles utilisent de nombreuses techniques. L'objectif est d'améliorer la qualité des résultats de recommandation en utilisant les avantages de différentes techniques tout en surmontant les inconvénients. Le problème important des méthodes hybrides est la combinaison efficace de techniques.

4.6. Comparaison entre les techniques de recommandation

Après avoir introduits les principales techniques de recommandation appliquées à la recommandation des papiers de recherche : filtrage à base de contenu, filtrage collaboratif et la méthode basée graphes. Le tableau 2 illustre les principaux avantages et les inconvénients des CBF, CF et GM.

- Chaque technique de recommandation peut surmonter les inconvénients d'autres techniques.
- CF peut surmonter le problème de qualité des résultats de la recommandation, mais il a encore le problème de démarrage à froid et d'autres inconvénients.
- Pour combiner les avantages et éviter les inconvénients de ces techniques, Les méthodes hybrides sont utilisées. Ces méthodes hybrides utilise CBF et CF pour faire des systèmes de recommandation plus efficaces, en plus, CBF et GB sont utilisés pour recommander des papiers.

Techniques	Avantages	Inconvénients
Filtrage a base du contenu (CBF)	• chaque article peut être découvert pour calculer la similarité • les résultats sont liés aux préférences personnelles des utilisateurs	• ne considérer que la pertinence des mots qualité est incertain • problème d'utilisateur nouveau
Filtrage collaboratif(CF)	• les résultats de la recommandation peuvent être fortuits • la qualité des résultats peut être garantie	• problème de démarrage à froid • problème de parcimonie
Méthode basée sur les graphes (GB)	• considère des sources différentes à recommander	• ne tient pas compte du contenu des articles et les intérêts des utilisateurs.

Tableau 2: Comparaison entre les techniques de recommandation.

5. Evaluations des systèmes de recommandation

Evaluer un système de recommandation permet de mesurer ses performances vis-à-vis de ses objectifs. De ce fait, les dimensions à évaluer diffèrent selon les objectifs fixés (prédiction des notes des utilisateurs pour les items, augmentation des ventes, etc.)

Selon [Herlocker et al. 2004], les évaluations des systèmes de recommandions peuvent être effectuées en utilisant une analyse hors ligne (offline analysis) ou une expérimentation avec des utilisateurs réels (live user expriment). Il existe aussi d'autre classification des méthodes

d'évaluation des systèmes de recommandation. Ces méthodes d'évaluation sont classées en trois types : expérimentations offline, études avec des utilisateurs (user studies) et tests réels (real life testing nommé aussi expérimentation en ligne (Online expriments).

5.1 Evaluation Online

L'évaluation online peut aussi recueillir le point de vue de l'utilisateur concernant le système de recommandation. Dans ce type d'évaluation, des utilisateurs réels utilisent le système dans des conditions réelles sur une longue période. Ce type d'évaluation peut montrer les usages et les habitudes de manipulation des utilisateurs, les problèmes et les besoins non satisfaits, et les problèmes que les chercheurs n'ont peut-être pas envisagés. Avec ces tests réels sur le terrain, la plupart des objectifs centrés sur l'utilisateur peuvent être efficacement évalués, comme l'évaluation de l'expérience utilisateur, la satisfaction des utilisateurs ou la rétention des utilisateurs [Herlocker *et al*, 2004] .

5.2 Evaluation Offline

Une grande partie du travail d'évaluation des algorithmes des systèmes de recommandation s'est concentrée sur l'analyse hors ligne de la précision des prédictions que peuvent faire ces systèmes. Les évaluations hors lignes utilisent des ensembles de données (dataset) constitués d'actions des utilisateurs. Les évaluations offlines simulent le processus de recommandation où un sous ensemble des actions utilisateurs du dataset est caché et le système de recommandation prédit ces actions cachées. Le système de recommandation est évalué en fonction de sa capacité à prédire ces interactions cachées. Les résultats de ces prédictions sont analysés en utilisant une ou plusieurs métriques. Deux types d'ensembles de données sont souvent utilisés dans ces évaluations [Erdt et al. 2015] :

- Ensembles de données naturels : ils sont constitués de données issues de l'historique des interactions d'utilisateurs réels dans un système donné sur une période donnée. De nombreux dataset sont disponibles pour mener des évaluations sur des algorithmes de recommandation.
- Ensembles de données de synthèse : ils sont construits de données artificielles. Ce type de dataset est habituellement utilisé pour tester comment les algorithmes de recommandation fonctionnent dans certaines conditions [Tadlaoui, 2018].

Les évaluations offlines ont l'avantage d'être rapide et économique et peuvent être réalisées sur plusieurs ensembles de données ou plusieurs algorithmes différents à la fois. Ce type d'évaluation est une évaluation objective des résultats de la prédiction. Aucune analyse hors ligne ne peut déterminer si les utilisateurs préfèrent un système particulier, soit en raison de ses

prédictions, soit en raison d'autres critères moins objectifs tels que l'esthétique ou l'ergonomie de l'interface utilisateur [Tadlaoui 2018].

5.2.1 Protocoles d'évaluation

Une fois le dataset préparé, un protocole d'évaluation doit être sélectionné, Il en existe trois [Tadlaoui, 2018] :

- **Retrait (Hold out) :** un pourcentage du dataset est utilisé comme données d'apprentissage tandis que le reste est utilisé comme données de test.
- **Tous sauf 1 (All but 1) :** Un seul élément du dataset est utilisé comme données de test, ce qui maximise la quantité de données utilisées comme données d'apprentissage.
- **Validation croisée K-fold (K-fold cross validation) :** Les données d'évaluation sont divisées K fois en données de formation et en données de test.

Pour chacune des k validations croisées, il est possible d'utiliser le retrait ou le tout sauf un. La validation croisée K-fold est plus intéressante que les deux premières car elle permet de maximiser la possibilité qu'un élément du dataset soit dans les données de tests et les données d'apprentissage.

5.2.2 Métriques d'évaluation

La performance des systèmes de recommandation est testée en termes de la précision des listes de recommandation générées. Les mesures d'évaluation de la performance des recommandations Top-N sont : la Précision, le Rappel et la métrique F1 [Isinkaye et al 2015].

- **La précision :** est une mesure d'exactitude. Elle détermine la proportion des items pertinents recommandés parmi tous les items recommandés : autrement dit, la proportion des recommandations qui se sont avérées pertinentes.

Elle est définie par : $$\mathbf{P}=\frac{N_t}{N}$$

-**Nt** : Nombre des items pertinents trouvés.

- **N** : Nombre total des items.

- **Le rappel :** est une mesure d'exhaustivité qui détermine la proportion des items pertinents recommandés parmi tous les items pertinents .

Autrement dit, c'est le rapport entre le nombre d'items pertinents sélectionnés et le nombre total des items pertinents disponibles. Ceci, représente la probabilité qu'un item pertinent soit sélectionné.

$$\mathbf{R}=\frac{N_t}{N_p}$$

-$\mathbf{N_t}$ **:** Nombre des items pertinents trouvés

-$\boldsymbol{N_p}$: Nombre total des items pertinents sélectionnés.

Les deux mesures citées ci-dessus (Précision et Rappel) ne mesurent qu'une pertinence binaire (i.e. item pertinent/non pertinent), c'est-à-dire elles ne peuvent pas mesurer la qualité d'ordonnancement des items pertinents sélectionnés [Maatallah, 2015].

- **La métrique F1 :** C'est une combinaison des deux métriques précédentes, et elle est définie par : $$\boldsymbol{F_1 = \frac{2PR}{P+R}}$$

D'autres métriques sont utilisées pour évaluer les systèmes de recommandation telle l'erreur quadratique absolue (MAE) et l'erreur quadratique moyenne (RMSE) [Isinkaye et al 2015].

- MAE : calcule l'écart entre les notes prédites et les vraies notes.
- RMSE : elle repose sur le même principe que la MAE, tout en mettant l'accent sur les écarts importants.

6. Conclusion

Ce chapitre a présenté un certain nombre d'approches visant à produire des systèmes de recommandation qui se catégorisent principalement en approches basées sur le contenu, et les approches utilisant un filtrage collaboratif.

Nous avons évoqué aussi l'utilisation de ces approches pour la recommandation des papiers de recherches. Dans ce domaine nous avons également souligné l'utilisation de l'approche basée sur les graphes dans plusieurs travaux, et sur laquelle repose notre travail. Des approches d'hybridation des techniques précédentes ont été proposées pour améliorer les résultats de la recommandation.

Chapitre 2 : Conception du système proposé pour la recommandation des papiers de recherches

1. Introduction

L'utilisation des moteurs de recherche génériques lors de la recherche d'informations connexes sur Internet est devenue la méthode la plus courante et la plus pratique. Cependant, les résultats de cette approche dépendent largement de la capacité de l'utilisateur à affiner le message de requête afin de personnaliser les résultats de la recherche.

Une autre approche classique utilisée par la plupart des chercheurs consiste à suivre la liste des références des documents qu'ils possédaient déjà [Sun *et al* ,2014]. Même si cette approche peut être assez efficace dans certains cas, elle ne garantit pas une couverture complète des articles scientifiques et ne peut pas retracer les articles publiés après l'article en possession. De plus, la liste des références peut ne pas être accessible au public et donc difficile d'accès pour les chercheurs.

Une approche alternative qui a été proposée dans la littérature est l'utilisation de systèmes de recommandation d'articles scientifiques, pour suggérer automatiquement des articles pertinents aux chercheurs sur la base de certaines informations initiales qui sont plus élaborées que quelques mots clés. Dans notre travail après avoir revue les différentes techniques de recommandation des papiers de recherche nous avons opté pour une approche collaborative reposant sur le réseau des citations.

En fait, différents chercheurs ont proposé une utilisation différente des informations fournies par les utilisateurs. Pour être précis, a exploré l'utilisation de l'approche de filtrage collaboratif pour recommander des articles scientifiques à un chercheur à partir de l'ensemble des citations vers l'un de ses articles. L'objectif était de tester la capacité de l'approche de filtrage collaboratif à recommander un ensemble de citations qui seraient très importantes en tant que références supplémentaires à un document cible.

Le réseau de citations a été exploré, pour améliorer les performances de la recommandation. Cependant, l'approche génère un problème de rareté dans le réseau de citation papier et rend ainsi le processus de recommandation très délicat. Pour atténuer le problème de rareté, [Aznoli et Navimipour, 2016] a appliqué le concept de l'approche de filtrage collaboratif pour identifier les articles de citation potentiels à partir de la liste des articles rédigés par un

chercheur. Les résultats expérimentaux montrent que les recommandations après avoir découvert les articles de citation potentiels sont plus efficaces que le filtrage collaboratif avec des valeurs binaires ou de similarité. Sous cet aspect nous avons fondé l'approche présentée dans ce mémoire.

2. Présentation générale de l'approche proposée

Motivé par les travaux [Aznoli et Navimipour, 2016] [Haruna et al, 2018]. Notre travail consiste à la proposition d'une approche collaborative pour la recommandation d'articles de recherche. L'approche est basée sur un réseau de citations qui en plus d'explorer les associations entre un papier cible, ses références et ses citations, permet également de mettre en évidence les papiers en relation avec les citations du papier cible, ainsi qu'avec ses références. Notre objectif est d'exploiter le potentiel des associations qui existent entre les articles de recherche en se basant sur la perspective des relations papier-citation. Nous mesurons et évaluons ensuite l'étendue de la similitude entre le papier et les papiers candidats par le calcul de fréquence et des poids de similarité et nous recommandons les N-papiers les plus similaires au papier cible.

L'approche proposée illustrée par la figure 8 s'articule autour de quatre étapes principales à savoir :

- Extraction des papiers en relation avec le papier cible ;
- Le calcul des fréquences des citations ;
- Calcul des similarités ;
- Classements des papiers candidats.
- Affichage de la liste des recommandations

Avec ces étapes un prétraitement est effectué pour transformer le corpus en matrice de citations.

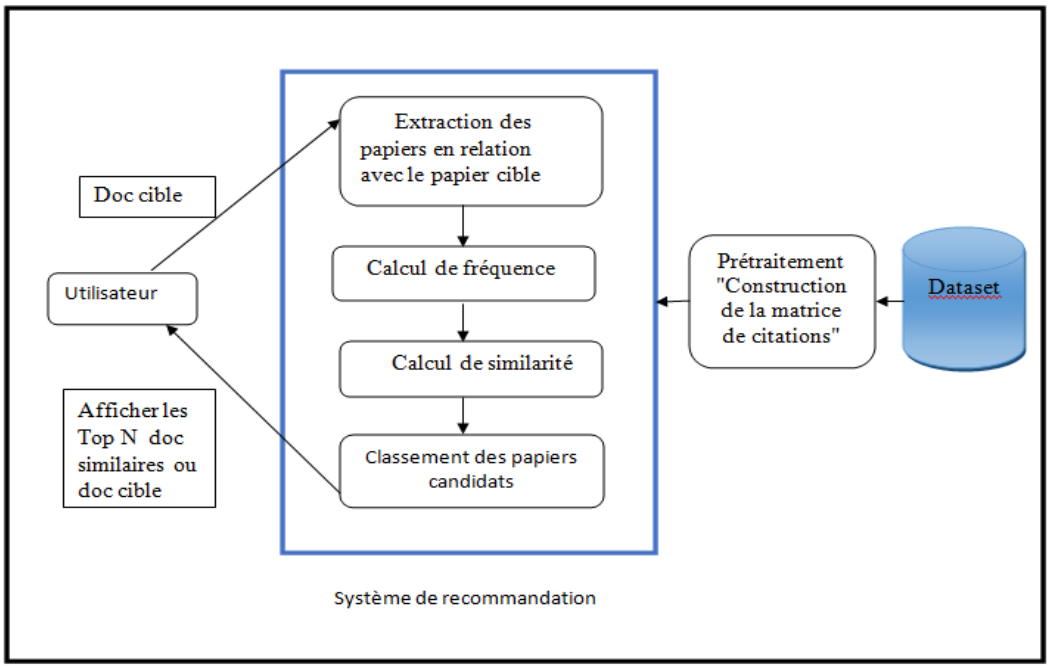

Figure 8 : Architecture générale de l'approche proposée

3. Conception détaillée de l'approche proposé

L'approche proposée commence par la transformation du corpus en une matrice de citation des papiers (Mcit). Le papier cible (P_i) est défini comme étant le papier dont un chercheur possède et souhaite recevoir d'autres recommandations qui lui sont similaires.

À la réception de la requête de l'utilisateur, l'approche proposée identifie le papier cible dans la matrice de citations des papiers et puis l'algorithme 1 est appliqué.

Algorithme 1

Algorithme : Un système de recommandation à base de citation

Entrée : Papier Cible

Sortie : Recommandation Top -N

Etant donné un papier cible Pi comme requête , construire la liste des papiers candidats comme suit :

1)- Extraire de la matrice des citations, l'ensemble des papiers directement reliés au papier Pi : ses citations (Cit j) et ses références Ref j.

2)- Pour chaque citation Citj récupérer ses citations et ses références et pour chaque papier Ref j récupérer ses citations et ses références.

le papier cible sont collectés de la matrice Mcit et constituent les papiers candidats à la recommandation. Ensuite, la similarité est mesurée entre le papier cible et chacun des papiers candidats. La similarité est calculée en utilisant la fréquence de chaque papier en considérant le nombre de ses citations et ses références. Pour donner une importance aux documents ayant une relation directe avec le papier cible, leurs fréquences sont multipliées par un coefficient. Enfin, les poids de similarité des papiers candidats sont calculés par la normalisation de la fréquence en la divisant par le maximum des fréquences augmenté de 1. De ce fait, les poids prennent des valeurs dans l'intervalle [0,1]. Généralement, les résultats ne sont pas ordonnés, alors nous procédons à un classement les papiers candidats selon leurs poids de similarité et nous choisissons les N meilleurs papiers ayant les poids de similarité les plus élevés. Finalement, l'algorithme recommande les N papiers les plus similaires au papier cible à l'utilisateur.

Pour comprendre clairement l'approche proposée, la Figure 9 représente un document cible (p_i) avec des citations (Cit.1 à Cit.N), dans lequel chacune des citations a référencé un ensemble d'autres articles (Rec.1 à Rec. N).

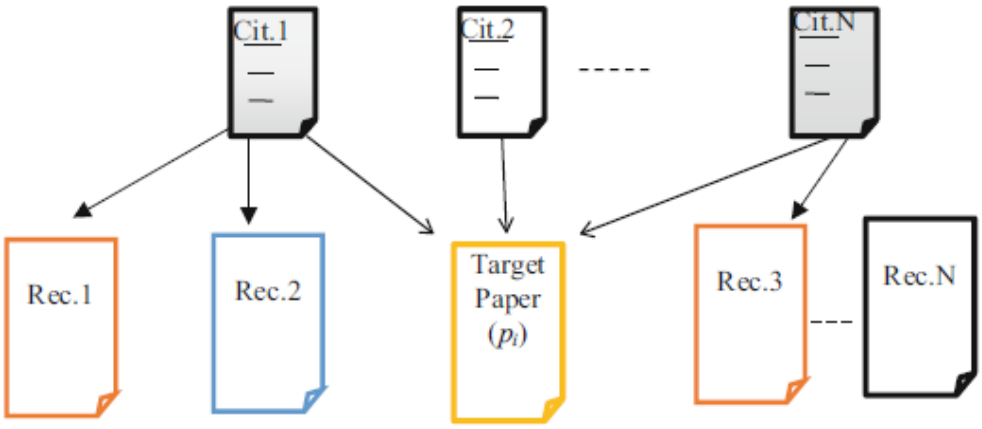

Figure 9 : Relation du papier cible avec les papiers co-référencés.

Le but est de mesurer l'étendue de la similarité entre le papier cible (P_i) et chacun des papiers coréférences. Pour ce faire, les relations contextuelles entre le papier cible et ses papiers voisins sont exploitées pour transformer la matrice de citation papier en une matrice relationnelle pour représenter le papier cible (P_i) concernant chacun de ses papiers voisins.

L'exemple de la figure 10 représente la construction de la matrice de citation, puis en figure 11 la matrice relationnelle obtenue suite au calcul des fréquences de citations .

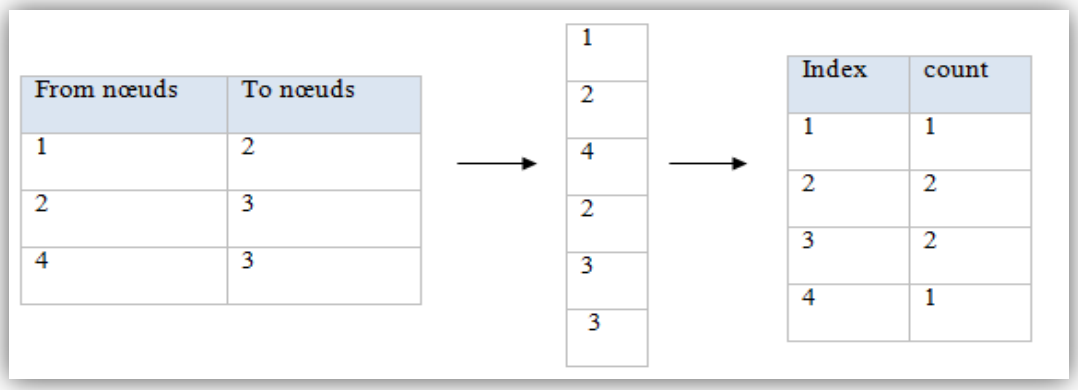

Figure 10 : Matrice de citation.

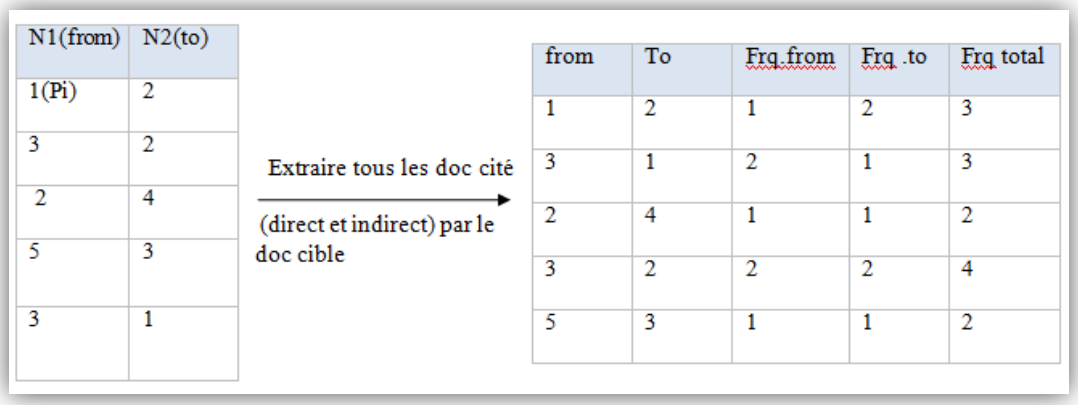

Figure 11: Matrice relationnelle

4. Conclusion

Dans ce deuxième chapitre, nous avons présenté l'approche proposée et son architecture et nous avons présenté aussi d'une manière détaillée la conception de notre système de recommandation des papiers de recherches. Le prochain chapitre présente l'implémentation et le fonctionnement de notre application.

Chapitre 3 : Implémentation

1. Introduction

Dans cette section, nous présentons les outils de développement, le détail de l'implémentation du prototype de l'application, avec son expérimentation sur des données réelles ainsi que la description des différentes fonctionnalités à travers les interfaces de l'application.

2. Outils et environnement de développement

Deux types de plateformes sont à présenter ici, au niveau de la plateforme matérielle, on va présenter la machine auquel on a réalisé et tester notre système, avec une description de la configuration matérielle de l'ordinateur utilisé pendant le développement. Une plateforme logicielle représente les outils et les langages de programmation.

2.1 Plateforme matérielle

Pour réaliser notre système, on a utilisé la configuration matérielle suivante :

PC Acer :Processeur (Intel® Core™ i5-4200U CPU@ 2.6GHz), RAM (4.00 Go).

Système d'exploitation : Windows 7 Professionnels 64 bits.

2.2 Plateforme logicielle

Dans cette partie nous exposerons brièvement l'environnement de développement. Nous évoquerons les outils, les langages de programmation et les utilitaires.

A. Langages de programmation

- **Python** est le langage de programmation open source le plus employé par les informaticiens. [Net 01]

Ce langage s'est propulsé en tête de la gestion d'infrastructure, d'analyse de données ou dans le domaine du développement de logiciels. En effet, parmi ses qualités, Python permet notamment aux développeurs de se concentrer sur ce qu'ils font plutôt que sur la manière dont ils le font. Il a libéré les développeurs des contraintes de formes qui occupaient leur temps avec les langages plus anciens. Ainsi, développer du code avec Python est plus rapide qu'avec d'autres langages. Les principales utilisations de Python par les développeurs sont :

- la programmation d'applications
- la création de services web
- la génération de code
- la méta-programmation

B. Utilitaires

➢ **Jupyter** :

est une application web utilisée pour programmer dans plus de 40 langages de programmation, dont Python, Julia, Ruby, R, ou encore Scala . Jupyter est une évolution du projet IPython. il permet de réaliser des calepins ou notebooks, c'est-à-dire des programmes contenant à la fois du texte en markdown et du code en Julia, Python, R. Ces calepins sont utilisés en science des données pour explorer et analyser des données [Net 02].

➢ **Kaggle**

est une plateforme web organisant des compétitions en science des données. Sur cette plateforme, les entreprises proposent des problèmes en science des données et offrent un prix aux datalogistes obtenant les meilleures performances. L'entreprise a été fondée en 2010 par Anthony Goldbloom. [Net 03]

➢ **Visual Studio Code**

Est un éditeur de code source gratuit créé par Microsoft pour Windows, Linux et macOS. Les fonctionnalités incluent la prise en charge du débogage, de la coloration syntaxique, de l'achèvement de code intelligent, des extraits de code, du refactoring de code et de Git intégré. Les utilisateurs peuvent modifier le thème, les raccourcis clavier, les préférences et installer des extensions qui ajoutent des fonctionnalités supplémentaires.

Le code source de Visual Studio Code provient du projet VSCode de logiciel libre et open source de Microsoft publié sous la licence expat permissive, et les binaires compilés sont des logiciels gratuits pour toute utilisation [Net04].

3. Expérimentation

3.1 Dataset

Nous avons utilisé une dataset fournie par la Coupe KDD 2003, dans nos expériences: le HEP-PH (phénoménologie de la physique des hautes énergies). Cette dataset a été extraite de site Web : l'e-print arXvi.org et incluent 205406 relations de citation, 10760 articles cités et 6580 articles de citation. Les distributions de l'ensemble de données prétraité sont présentées dans le tableau 3. Comme la plupart des ensembles de données dans le domaine des systèmes de recommandation, les relations de citations dans l'ensemble de données sont très rares (0.9971 pour Hep-PH), c.-à-d. la rareté des données. La rareté indique le rapport de la différence entre les nombres de toutes les relations possibles et les relations de citation existantes au nombre de toutes les relations possibles.

Dataset	HEP-PH
nombre d'articles de citation	6580
nombre d'articles cités	10760
nombre de relations de citation	205406
rareté des relations de citation	0.9971

Tableau 3 : Statistiques des données.

3.2. Description des fonctionnalités

Dans cette partie, nous allons présenter un aperçu sur quelques fonctionnalités de notre système de recommandation à travers ses interfaces.

A. Interface principale

Cette interface permet à un utilisateur de charger la base de données et d'introduire une requête en spécifiant un papier cible.

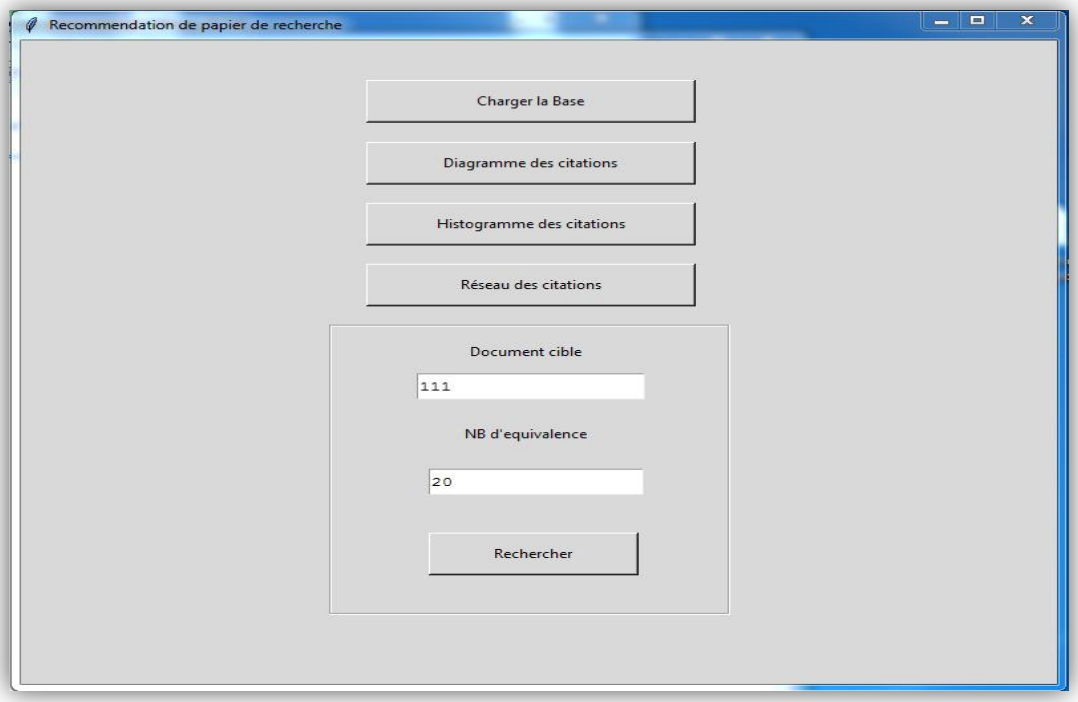

Figure 12 : L'interface de notre système de recommandation des papiers scientifiques.

B. Réseau de citations

La figure 13 représente le réseau de citations du corpus présent dans le dataset cit-Hep.

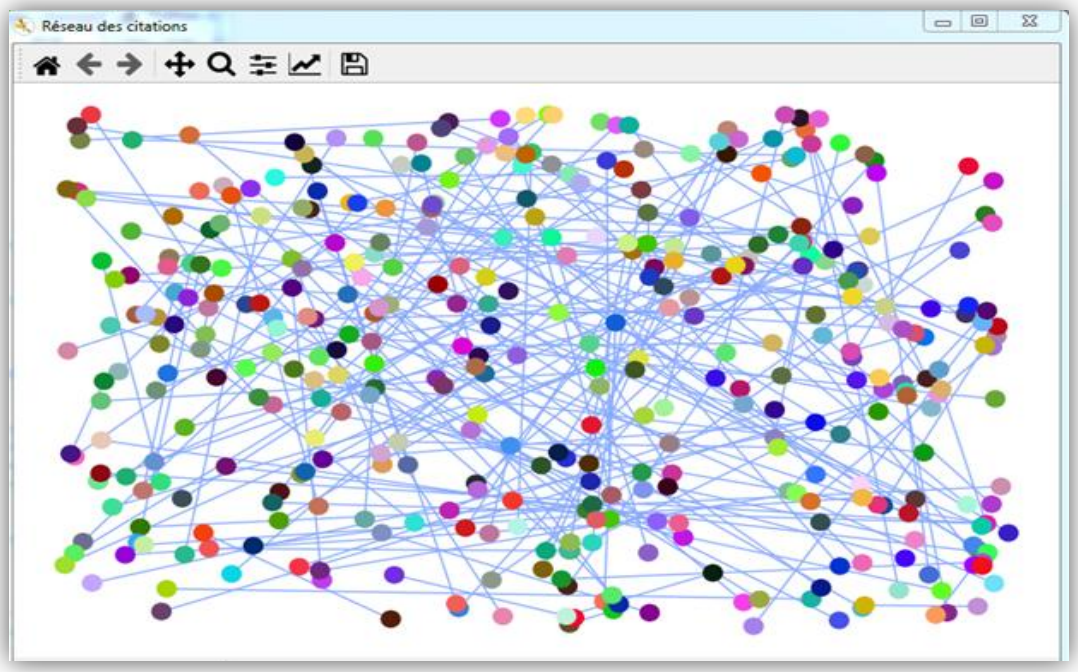

Figure 13 : Réseau de citations.

C. Diagramme de citations

La fenêtre suivante affiche le réseau de citations précédant en termes de nombre de citations (count).

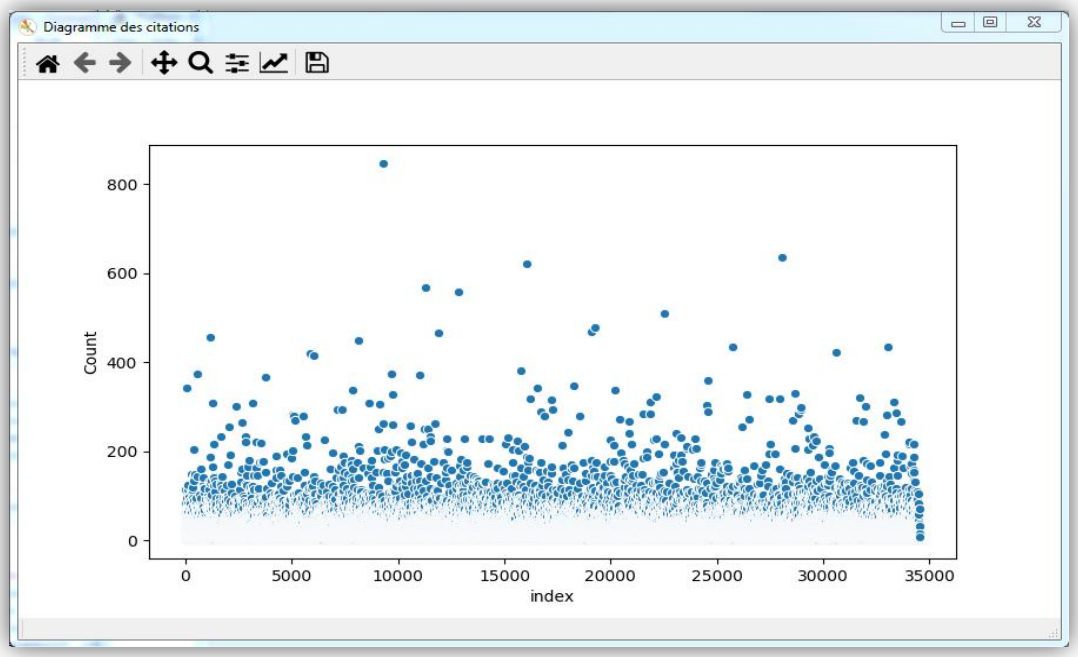

Figure14 : Diagramme de citations

D. Histogramme de citations

L'histogramme représente les fréquences des documents indexé

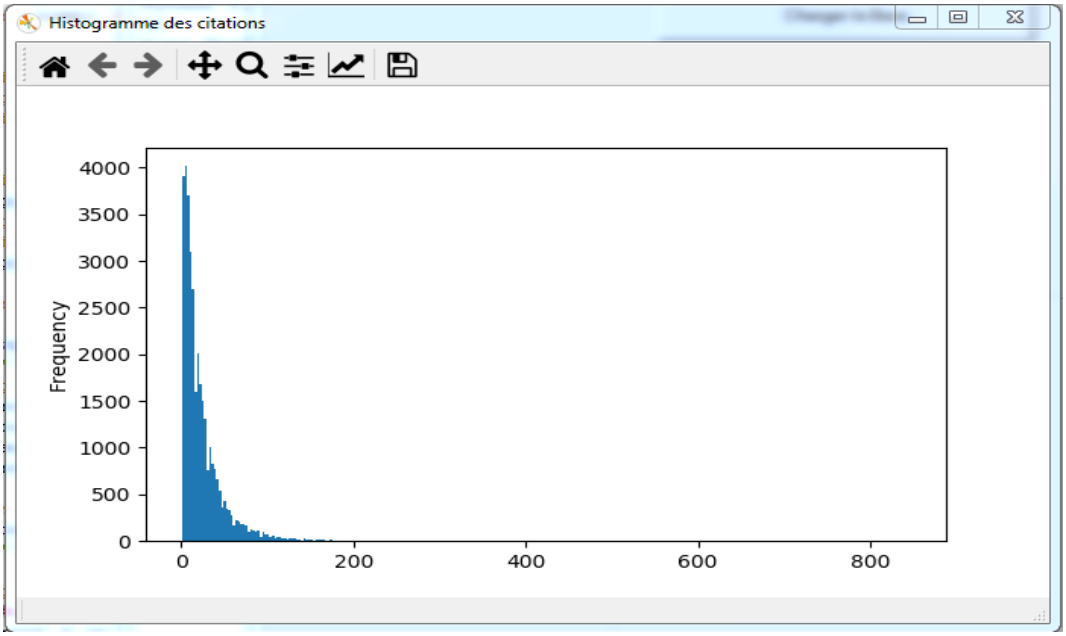

Figure15 : Histogramme de citations.

E. Recommandation

- **Calcul de similarité**

Un fois la requête utilisateur saisie. La procédure de calcul de similarité est exécutée et les résultats sont visualisés sous forme de diagrammes (figure 16 et 17).

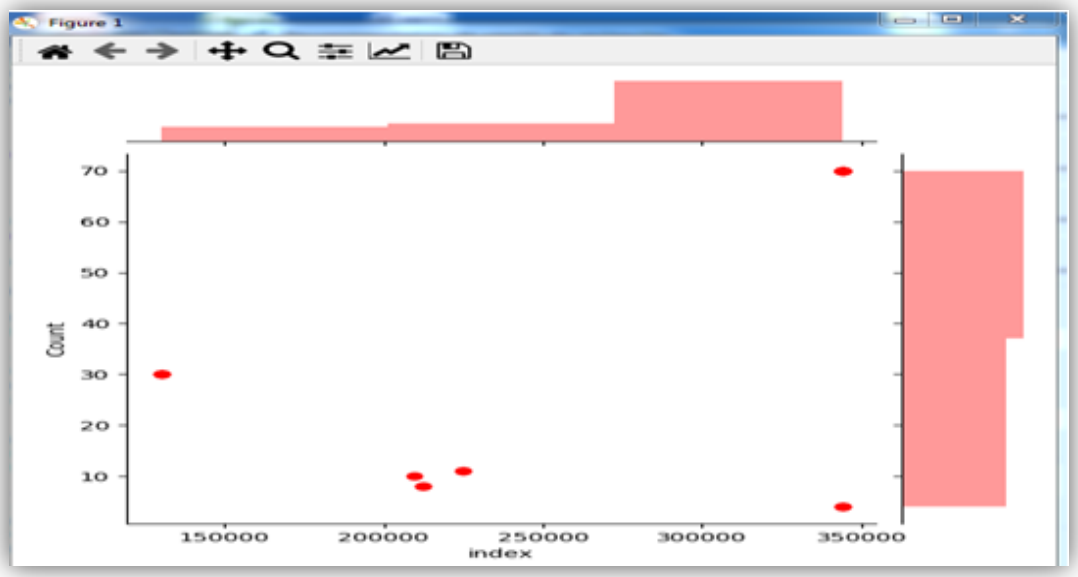

Figure 16 : Diagramme des nœuds similaire.

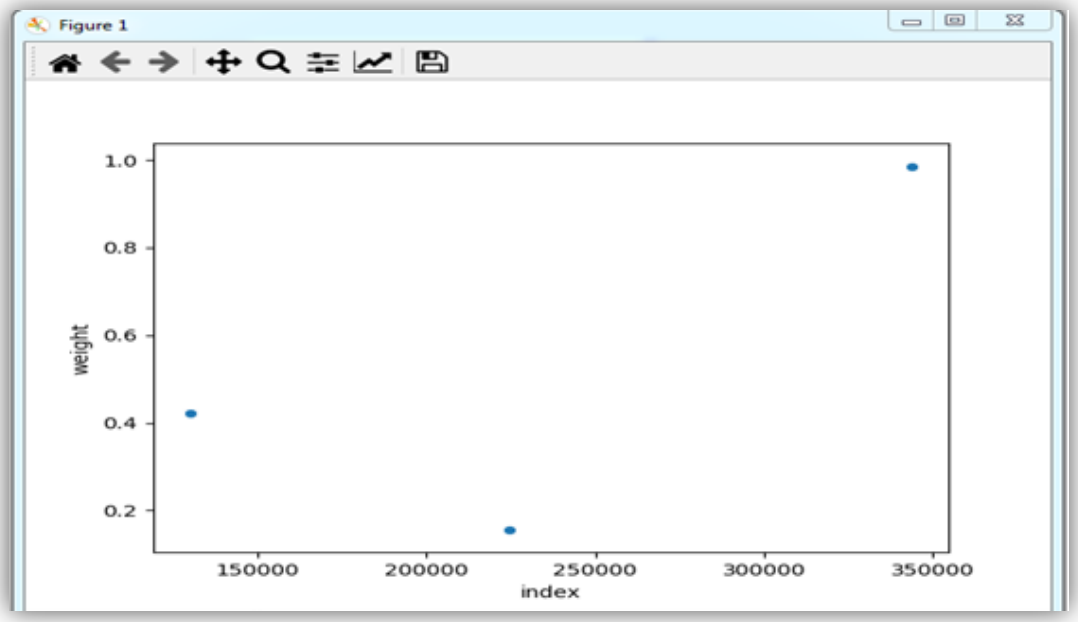

Figure 17 : Diagramme de poids des papiers similaires.

- **Résultat de la recommandation**

Enfin le système nous affiche une simple arbre qui représente les 20 top documents similaires au document cible proposé par l'utilisateur avec le poids de chaque document similaire, Habituellement, le système affiche 20 documents similaires, mais dans le cas où le système affiche moins de 20 documents similaires, cela signifie que le document cible a moins de 20 documents similaires et ceci est basé sur le poids des documents calculé par le système (figure 18).

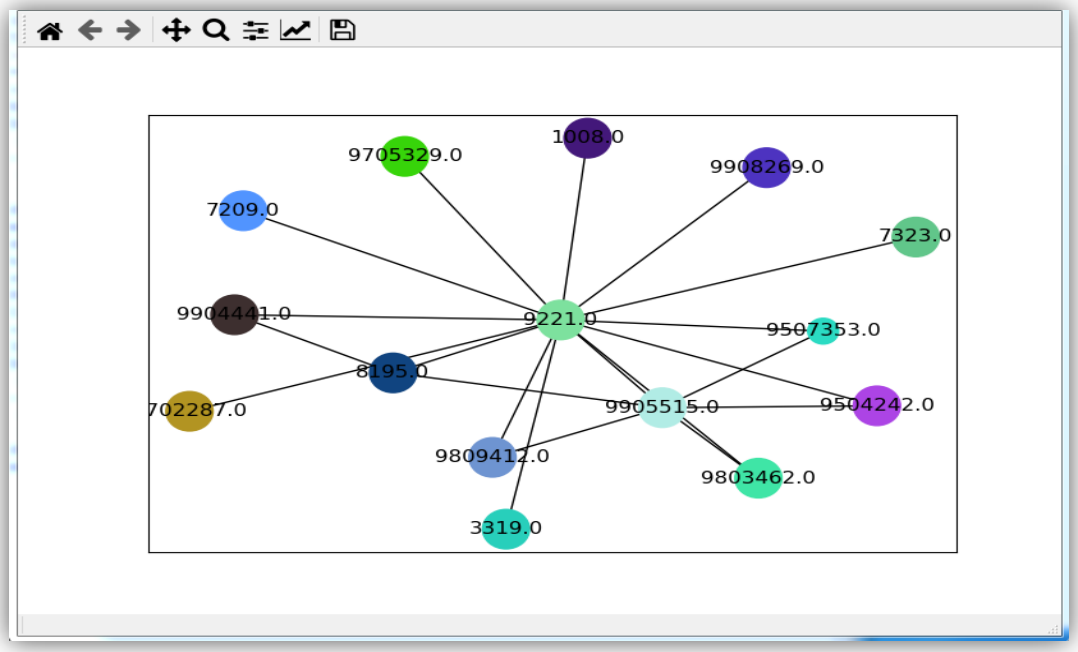

Figure 18 : l'arbre des 20 top documents similaires.

4. Extraits de codes

Des parties de codes de l'application développée sont présentées dans la figure 19.

In [ ]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import networkx as nx
import random
%matplotlib inline
```

In [ ]:

```
f = open('data.txt')
s = f.readline()
data=[]
while s!='':
    s = s.replace('\n','')
    values=s.split('\t')
    data.append(tuple((int(values[0]),int(values[1]))))
    s = f.readline()
df = pd.DataFrame(data,columns=['From Node','To Node'])
#print(f'Nombre total de relation : {len(data)}')
df.info()
```

In [ ]:

```
f = open('data.txt')
s = f.readline()
data=[]
while s!='':
    s = s.replace('\n','')
    values=s.split('\t')
    data.append(tuple((int(values[0]),int(values[1]))))
    s = f.readline()
df = pd.DataFrame(data,columns=['From Node','To Node'])
#print(f'Nombre total de relation : {len(data)}')
df.info()
```

In [ ]:

```
....
```

In [ ]:

```
plt.figure(figsize=(12,5))
sample_count=200
g = nx.from_pandas_edgelist(df.sample(sample_count),'From Node','To Node')
nx.draw_random(g,with_labels=False,node_size=100,node_color=randomColor(len(g.nodes)),e
dge_color='#88aaff')
```

```
total_cites=total_cites.join(total_cites['From Node'].value_counts(),on='From Node',rsu
ffix=' cnt')
tmp = df['To Node']
tmp.rename('From Node',axis=1,inplace=True)
total_cites=total_cites.join(tmp.value_counts(),on='From Node',rsuffix=' cntright')
total_cites['From Node cntright'].rename_axis('cntt',inplace=True)
total_cites.fillna(0,inplace=True)
total_cites['Count']=total_cites['From Node cnt']+total_cites['From Node cntright']
total_cites.drop(axis=1,columns=['From Node cnt','From Node cntright'],inplace=True)
```

Figure 19: Extraits des codes de l'application**.**

4. Conclusion

Dans ce chapitre, nous avons présenté le détail de l'implémentation de notre application de recommandation des papiers de recherche.

Nous avons aussi décrit le détail de l'expérimentation que nous avons effectué.

Conclusion et perspectives

L'utilisation de systèmes de recommandation pour extraire des papiers de recherche pertinents est devenue vitale en raison du défi récent de la gestion des données de recherche massives car ces systèmes jouent un rôle important dans la recherche et le filtrage des informations. Dans ce mémoire nous avons étudié un ensemble d'approches pour la recommandation des papiers de recherche. En fait, ces approches reposent sur différentes techniques, en particulier : les techniques basées contenu, les techniques de filtrage collaboratif et les techniques à base de graphes. Ainsi, nous avons proposé une approche collaborative à base de graphe pour la recommandation des papiers de recherches. Le principe de notre approche est étant donné un papier d'intérêt, il s'agit de construire un réseau de citations qui représente le papier et ses relations directes avec les papiers référencés et cités et des relations indirectes de co-citation ainsi qu'avec les papiers citant le papier cible. L'approche développée est collaborative, le filtrage est implicitement collaboratif entre les papiers dans le sens où les papiers référenciés dans un papier particulier sont considérés comme évalués positivement par ce papier. Pour la concrétisation de ces aspects un prototype de l'application est développé et expérimenté. Néanmoins, l'évaluation prévue de ce prototype n'est pas réalisée.

En perspectives, ce travail permet d'ouvrir plusieurs voies d'amélioration et de développement. Initialement, l'évaluation de cette application est primordiale. En second lieu, l'ajout au niveau du réseau des citations d'une couche pour la prise en compte des auteurs et les relations auteur-papier est très intéressante et apportera une valeur ajoutée à la qualité de la recommandation. Une autre alternative aussi est l'exploration de la possibilité d'hybridation de cette approche avec une méthode basée sur le contenu des papiers en particulier les titres et les résumés s'avère bénéfique et influencera significativement l'efficacité et la performance de la recommandation. Enfin, l'ajout d'une couche sémantique à notre système en utilisant les ontologies et les techniques de web sémantique est également un plus à étudier.

Références

[Burke, 2002]

Hybrid Recommender Systems: Survey and Experiments† Robin Burke Grégoire.

[Adomavicius et al, 2005]

Adomavicius, G.,& Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering, 17(6), 734-749.

[Lops *et al*, 2011]

Pasquale Lops, Marco de Gemmis and Giovanni Semeraro. Content-based Recommender Systems: State of the Art and Trends chapitre dans le livre "Recommender Systems Handbook". Éditeurs Francesco Ricci · Lior Rokach · Bracha Shapira· Paul B. Kantor

[Bechet, 2012]

Nicolas Béchet, Etat de l'art sur les Systèmes de Recommandation Projet AxIS de l'INRIA, dans le cadre du projet Addictrip Available . Disponible en ligne sur http://people.irisa.fr/Nicolas.Bechet/Publications/EtatArt.pdf

[Shardanand &al,95]

Shardanand, U. and Maes, P. (1995). Social Information Filtering: Algorithms for Automating "Word of Mouth". pp. 210-217, ACM Press. Disponible en ligne sur http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.449.5366&rep=rep1&type=pdf

[van Rijs et al , 79]

van Rijsbergen, C. and D, P. (1979). Information Retrieval.

[Vozalis et al 2003]

Vozalis, E. Margaritis, K.G. (2003). Analysis of Recommender Systems' Algorithms. 6th Hellenic European Conference on Computer Mathematics & its Applications (HERCMA), Athens, Greece.

[Herlocker et al 2004]

Herlocker, J.L., Konstan, J.A., Riedl, J.T. & Terveen, L.G. (2004). Evaluatingcollaborative filtering recommender systems. ACM Trans. Inf. Syst., 22(1), 5-53.

[Pazzani, 1999]

Pazzani, M.J. (1999). A Framework for Collaborative, Content-Based and Demographic Filtering. Artif. Intell. Rev., 13(5-6), 393-408.

[Naak 209]

Naak, A. (2009). Papyres : Un système de gestion et de recommandation d'articles de recherche. Mémoire présenté à la Faculté des études supérieures en vue de l'obtention du grade de Maîtrisées Sciences en Informatique, Montréal, Canada. Disponible en ligne https://core.ac.uk/download/pdf/55645772.pdf

[Arnautu , 2012]

Octavian Rolland Arnautu (2012). Mures : Un système de recommandation de musique.Mémoire présenté à la Faculté des arts et des sciences en vue de l'obtention du grade de Maîtriseès Sciences en Informatique, Montréal, Canada. Disponible en ligne sur https://docplayer.fr/2536196-Mures-un-systeme-de-recommandation-de-musique.html

[Rao et Talwer, 2008]

Rao,N. and Talwar,V : Application domain and functional classification of recommender systems a survey. Desidoc journal of library and information technology, 28(3) :17–36, (2008)

[Golberg et al 1992]

David Golberg, David Nichols Brian M. Oki, Douglas Terry, "Using collaborative filtering to weave an information tapestry". Dans Communications of the ACM, December 1992 , volume 35, issue 12

[Resnick et al 1994]

Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, John Riedl "GroupLens: An Open Architecture for Collaborative Filtering of Netnews" Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work, Chapel Hill, NC: Pages 175-186

[Ait Ahmed Nora,2017]

Ait Ahmed Nora , Idris Khodja Asma, «Système de Recommandation de Cours à Base d'Ontologie », Mémoire de master Université de saida (2017-2018). Disponible en ligne sur: https://pmb.univ- aida.dz/butecopac/doc_num.php?explnum_id=684

[Papagelis, 2005]

Papagelis, M. (2005) Alleviating the Sparsity Problem of Collaborative Filtering Using Trust Inferences. Proceedings of the Third International Conference on Trust Management,3477,224-239.Disponible en ligne : http://dx.doi.org/10.1007/11429760_16

[Tadlaoui, 2018]

TADLAOUI Mohammed , Système de recommandation de ressources pédagogiques fondé sur les liens sociaux : formalisation et évaluation, Thèse de doctorat de l'université de Tlemecen (2018) disponible en ligne http://dspace.univ-tlemcen.dz/bitstream/112/13027/1/Systeme-de-recommandation-de-ressources.pdf

[Erdt et al, 2015]

Erdt, M., Fernandez, A., & Rensing, C. (2015). Evaluating recommender systems for technology enhanced learning: a quantitative survey. IEEE Transactions on Learning Technologies, 8(4), (pp. 326-344).

[Maatallah, 2015]

Majda MAATALLAH,(2015) « Une Technique Hybride pour les Systèmes de Recommandation ». Thèse de doctorat de l'université de Annaba . Disponible en ligne: http://biblio.univ-annaba.dz/wp-isponible content/uploads/2016/11/These-Maatallah-Majda.pdf

[Isinkaye et al 2015]

F.O. Isinkaye a,*, Y.O. Folajimi b, B.A. Ojokoh Recommendation systems: Principles, methods And Evaluation. Egyptian Informatics Journal Egyptian Informatics Journal (2015) 16, 261–273

[Amy et al., 2013]

Amy J.C. Trappey a,n , Charles V. Trappey b , Chun-Yi Wu a , Chin Yuan Fan c , Yi Liang Lin Intelligent patent recommendation system for innovative design collaboration ,Journal of Network and Computer Applications. Disponible en ligne sur : https://ir.nctu.edu.tw/bitstream/11536/23469/1/000328523000005.pdf

[liu et al,2015]

H. Liu, Z. Yang, I. Lee, Z. Xu, S. Yu, and F. Xia, ‘‘CAR: Incorporating filtered citation relations for scientific article recommendation,’’ in Proc. IEEE Int. Conf. Smart City/SocialCom/SustainCom (SmartCity), Dec. 2015, pp. 513–518. Disponible en ligne sur: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8598708

[cao et al ,2017]

D.Cao et al., ‘‘Cross-platform app recommendation by jointly modeling ratings and texts,’’ACM Trans. Inf. Syst., vol. 35, no. 4, 2017, Art. no. 37. https://nextcenter.org/wp-content/uploads/2018/02/Cross-Platform-App-Recommendation-by-Jointly-Modeling-Ratings.pdf

[Sugiyama et al,2010]

Sugiyama and M.-Y. Kan,2010 , ‘‘Scholarly paper recommendation via user’s recent research interests,’’ in Proc. 10th ACM Annu. Joint Conf. Digit. Libraries, 2010, pp. 29–38. Disponible en ligne sur : https://www.researchgate.net/publication/220923970_Serendipitous_recommendation_for_scholarly_papers_considering_relations_among_researchers

[Bai et al, 2019]

XIAOMEI BAI , MENGYANG WANG , IVAN LEE ,ZHUO YANG , XIANGJIE KONG ,Scientific Paper Recommendation: A Survey. Disponible en ligne : file:///C:/Users/malek/Downloads/[Bai%202019]%20Scientific%20paper%20recommendation%20_A%20%20survey.pdf

[Sun et al ,2014]

J. Sun, J. Ma, Z. Liu, and Y. Miao , ‘‘Leveraging content and connections for scientific article recommendation in social computing contexts,’’ Comput. J., vol. 57, no. 9, pp. 1331–1342, 2014. http://thealphalab.org/papers/ScientificArticleRecommendationExploitingCommonAuthorRelationsandHistoricalPreferences.pdf

[Sharma et al,2017]

R. Sharma, D. Gopalani, and Y. Meena , ‘‘Concept-based approach for rese arch paper recommendation,’’ Mach. Intell. Cham, Switzerland: Springer, 2017, pp. 687–692. Disponible en ligne : https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8434216

[Jomsri et al ,2010]

P. Jomsri,S. Sanguansintukul, and W. Choochaiwattana, ‘‘A framework for tag-based research paper recommender system: An ir approach,’’ IEEE 24th Int. Conf. Adv. Inf. Netw. Appl. Workshops (WAINA), 2010, pp. 103–108. Disponible en ligne sur : https://www.scss.tcd.ie/joeran.beel/pubs/research_paper_recommender_system_evaluation--quantitative_literature_survey.pdf

[Basuet al,2012]

C. Basu, H. Hirsh, W. W. Cohen, and C. Nevill-Manning ‘‘Technical paper recommendation: A study in combining multiple information sources,’’ J. Artif. Intell. Res., vol. 14, no. 1, pp. 231–252, 2012. Disponible en ligne : https://arxiv.org/pdf/1106.0248.pdf

[Chen et al,2007]

T. Chen, W.-L. Han, H.-D. Wang, Y.-X. Zhou, B. Xu, and B.-Y. Zang, ‘‘Content recommendation system based on private dynamic user profile,’’ in Proc. IEEE Int. Conf. Mach. Learn., vol. 4, Aug. 2007, pp. 2112–2118. Disponible en ligne : http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.471.4858&rep=rep1&type=pdf

[Kohar et Rana,2012]

M. Kohar and C. Rana, ‘‘Survey paper on recommendation system,’’ Int. J. Comput. Sci. Inf.Technol., vol. 3, no. 2, pp. 3460–3462, 2012 , Disponible en ligne : http://ijcsit.com/docs/Volume%203/Vol3Issue2/ijcsit2012030234.pdf

[Beel et al ,2016]

J. Beel, B. Gipp, S. Langer, and L. Breitinger, ‘‘Research-paper recommender systems:A literature survey,’’ Int. J. Digit. Libraries, vol. 17, no. 4, pp. 305–338, 2016. https://isg.beel.org/pubs/2016%20IJDL%20---

%20Research%20Paper%20Recommender%20Systems%20--

%20A%20Literature%20Survey%20(preprint).pdf

[Parra et Brusilovsky,2010]

D. Parra-Santander and P. Brusilovsky ,‘‘Improving collaborative filtering in social Tagging systems for the recommendation of scientific articles,’’in Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Intell. Agent Technol., Aug./Sep. 2010, pp. 136 142. Disponible en ligne sur : http://dparra.sitios.ing.uc.cl/old_site/Parra_Brusilovsky_WI_2010.pdf

[Sarwaret al,2001]

B. Sarwar, G. Karypis, J. Konstan, and J. Riedl ‘‘Item-based collaborative filtering recommendation algorithms,’’ in Proc. ACM 10th Int. Conf. World Wide Web, 2001, pp. 285–295. Disponible en ligne sur : https://grouplens.org/site-content/uploads/Item-Based-WWW-2001.pdf

[Huang et al,2002]

Z. Huang, W. Chung, T.-H. Ong, and H. Chen ‘‘A graph-based recommender system for digital library,’’ in Proc. 2nd ACM/IEEE-CS Joint Conf. Digit. Libraries,2002, pp 65–73. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.66.5744&rep=rep1&type=pdf

[Tian et Jing,2013]

G. Tian and L. Jing, ‘‘Recommending scientific articles using birelational graph-based iterativerwr,’’ in Proc. 7th ACM Conf. Recommender Syst., 2013, pp. 399–402. https://link.springer.com/article/10.1007/s10462-020-09819-4

[M. Ohta et al,2011]

M. Ohta, A. Takasu, and T. Hachiki, ‘‘Related paper recommendation to support online-browsing of research papers,’’ in Proc. IEEE 4th Int. Conf. Appl. Digit. Inf. Web Technol. (ICADIWT), Aug. 2011, pp. 130–136. Disponible en ligne : https://www.semanticscholar.org/paper/Related-paper-recommendation-to-support-of-research-

Ohta-Hachiki/b66167fe07d639bb143f84f15c5906387a4d86b7

[Fouss et al ,2007]

F. Fouss, A. Pirotte, J.-M. Renders, and M. Saerens, and M. Saerens, ‘‘Random walk computation of similarities between nodes of a graph with application to collaborative recommendation,’’ IEEE Trans. Knowl. Data Eng., vol. 19, no. 3, pp. 355-369, Mar. 2007. Disponible en ligne sur: http://outobox.cs.umn.edu/Random_Walks_Collaborative_Recommendation_Fouss.pdf

[Xu, et al,2016]

Z. Xu, H. Jiang, X. Kong, J. Kang, W. Wang, and F. Xia, ‘‘Cross-domain item recommendation based on user similarity,’’ Comput. Sci. Inf. Syst., vol. 13, no. 2, pp. 359–373, 2016. http://www.doiserbia.nb.rs/img/doi/1820-0214/2016/1820-021416000007Z.pdf

[Garfield,1972]

E. Garfield ‘‘Citation analysis as a tool in journal evaluation,’’ Science, vol. 178, no. 4060, pp. 471–479, 1972. http://www.garfield.library.upenn.edu/essays/V1p527y1962-73.pdf

[Haveliwala,2003]

H. Haveliwala ‘‘Topic-sensitive PageRank: A context-sensitive ranking algorithm for Web search,’’ IEEE Trans. Knowl. Data Eng., vol. 15, no. 4, pp. 784–796, Jul. 2003. https://www.researchgate.net/publication/3297186_Topic-sensitive_PageRank_A_context-sensitive_ranking_algorithm_for_Web_search

[Winotoet al,2012]

P. Winoto, T. Y. Tang, and G. I. McCalla, ‘‘Contexts in a paper recommendation system With collaborative filtering,’’ Int. Rev. Res. Open Distrib.Learn.,vol.13, no.5,pp. 56–75, 2012. Disponible en ligne : http://www.irrodl.org/index.php/irrodl/article/view/1243/2367

[Sugiyama et Kan,2013]

K. Sugiyama and M.-Y. Kan,, ‘‘Exploiting potential citation papers in scholarly paper recommendation,’’ in Proc. 13th ACM/IEEE-CS Joint Conf. Digit.Libraries,2013, pp. 153–162. Disponible en ligne : https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9035495

[Beel, et al,2017]

Beel, A. Aizawa, C. Breitinger, and B. Gipp, ‘‘Mr. DLib: Recommendations-as-a-service (RaaS)for academia,’’in Proc. ACM/IEEE Joint Conf. Digit. Libraries (JCDL),Jun.2017, pp. 1–2. Disponible en ligne sur : https://arxiv.org/ftp/arxiv/papers/1703/1703.09108.pdf

[Aznoli et Navimipour,2016]

F. Aznoli and N. J. Navimipour, ‘‘Cloud services recommendation: Reviewing the recentadvances and suggesting the future research directions,’’ J. Netw. Comput. Appl., vol. 77, pp. 73–86, Jan. 2016. Disponible en ligne : https://www.semanticscholar.org/paper/Cloud-services-recommendation%3A-Reviewing-the-recent-Aznoli-Navimipour/251f97cdc47fcd1e5cd4d0417cd1314276a529b5

[Haruna K et al., 2018]

khalid haruna, Maizatul Akmar Ismail , A. B. Baffa , Tutut Herawan ,Citation-Based Recommender System for Scholarly Paper Recommendation , Computational Science and Its Applications – ICCSA 2018 (pp.514-525) . Disponible en ligne sur : https://www.researchgate.net/publication/326150736_A_Citation Based_Recommender_System_for_Scholarly_Paper_Recommendation

[Beel *et al* , 2011]
Beel, J., Gipp, B., Langer, S., Genzmehr, M.: Docear: an academic literature suite for searching, organizing and creating academic literature. In: Proceedings of the 2011 Joint International Conference on Digital Libraries, JCDL'11, pp. 465–466 (2011)

[Bogers & Bosch, 2008]
Toine Bogers, Antal van den Bosch, " Recommending Scientific Articles Using CiteULike "dans Proceedings of the 2008 ACM Conference on Recommender Systems, RecSys 2008, Lausanne, Switzerland, October 23-25, 2008

[Steinert et al, 2015]
L. Steinert, I.-A. Chounta, and H. U. Hoppe, ``Where to begin? Using network analytics for the recommendation of scientific papers," Proceeding.
Int. Workshop Groupware. Switzerland: Springer,2015, pp. 124_139.

[Wang et al , 2016]
Q. Wang, W. Li, X. Zhang, and S. Lu, ``Academic paper recommendation based on community detection in citation-collaboration networks," Proceeding Web Conference Switzerland: Springer, 2016, pp. 124_136.

[Wang et al , 2016]
G. Wang, X. R. He, and C. I. Ishuga, ``HAR-SI: A novel hybrid article recommendation approach integrating with social information in scientific social network," Knowledge Based Systems , vol. 148, no. 15, pp. 85_99, 2018.

❖ Sites internet consulté

- [Net01] Définition de python
https://www.journaldunet.fr/web-tech/dictionnaire-du-webmastering/1445304-python-definition-et-utilisation-de-ce-langage-informatique/

- [Net02] Définition de jupyter
https://fr.wikipedia.org/wiki/Jupyter

- [Net03] Définition de Kaggle
https://fr.wikipedia.org/wiki/Kaggle

- [Net04] Définition de visuel studio code
https://fr.wikipedia.org/wiki/Visual_Studio_Code

Résumé

Le volume d'articles de recherche dans les référentiels numériques augmente. Cette croissance significative des référentiels fait qu'il est assez difficile pour les chercheurs d'obtenir des articles de recherche connexes en réponse à leurs requêtes. Le problème s'aggrave lorsqu'un chercheur ayant une connaissance insuffisante de la recherche d'articles scientifiques utilise ces référentiels.

Plusieurs approches ont été proposées pour aider les chercheurs à acquérir des articles pertinents et utiles à partir de l'énorme quantité d'informations disponibles sur Internet. Dans ce mémoire nous présentons une solution alternative au problème de recommandation des papiers de recherche. L'approche proposée est une approche collaborative exploitant un réseau de citations qui relie directement le papier d'intérêt avec ses citations et ses références et considère des relations indirectes illustrant les co-citations du papier cible. La recommandation des N meilleurs papiers se fait suite à un calcul des poids de similarité qui se base sur l'évaluation des fréquences de citations. Un prototype réalisant cette approche est implémenté et expérimenté avec des données réelles, nous a permis de constater que les résultats sont acceptables et l'approche est prometteuse.

Mots clés : *Recommandation, papier scientifique, similarité ,réseau des citations …*

Abstract

The volume of research articles in digital repositories is increasing. This significant growth in repositories makes it quite difficult for researchers to obtain related research articles in response to their queries.The problem becomes worse when a researcher with insufficient knowledge of researching scientific articles uses these repositories.

Several approaches have been proposed to help researchers acquire relevant and useful articles from the enormous amount of information available on the Internet. In this thesis we present an alternative solution to the problem of recommending research papers. The proposed approach is a collaborative approach exploiting a network of citations that directly links the paper of interest with its citations and references and considers indirect relationships illustrating the co-citations of the target paper. The recommendation of the N Top papers is made following a calculation of the similarity weights which is based on the evaluation of the citation frequencies. A prototype realizing this approach is implemented and tested with real data, has allowed us to see that the results are acceptable and the approach is promising.

Keywords: *Recommendation, scientific paper, similarity, citation network ….*

ملخّص:

يتزايد حجم المقالات البحثية في المستودعات الرقمية. هذا النمو الكبير في المستودعات يجعل من الصعب جدًا على الباحثين الحصول على مقالات بحثية ذات صلة ردًا على استفساراتهم. تزداد المشكلة سوءًا عندما يكون الباحث ليس لديه معرفة كافية ببحث المقالات العلمية في هذه المستودعات.

تم اقتراح العديد من الأساليب لمساعدة الباحثين في الحصول على مقالات مفيدة وذات صلة من الكم الهائل من المعلومات المتاحة على الإنترنت. نقدم في هذه الرسالة حلاً بديلاً لمشكلة التوصية بالأوراق البحثية. النهج المقترح هو نهج تعاوني يستغل شبكة من الاستشهادات التي تربط بشكل مباشر الورقة محل الاهتمام مع الاستشهادات والمراجع الخاصة بها وتعتبر العلاقات غير المباشرة التي توضح الاستشهادات المشتركة للورقة المستهدفة. تتم التوصية بأوراق البحث الأفضل بعد حساب أوزان التشابه التي تعتمد على تقييم ترددات الاقتباس. تم تنفيذ واختبار النموذج الأولي الذي يحقق هذا النهج باستخدام بيانات حقيقية ، مما سمح لنا برؤية ان النتائج مقبولة والنهج واعد.

كلمات مفتاحية

التوصية - أوراق البحث - شبكة من الاستشهادات -التشابه ...