جامعة سيدي محمد بن عبد الله بفاس
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵙⵉⴷⵉ ⵎⵓⵃⵎⴷ ⴱⵏ ⵄⴱⴷⵍⵍⴰⵀ ⴰ ⴼⴰⵙ
UNIVERSITÉ SIDI MOHAMED BEN ABDELLAH DE FES

كلية العلوم ظهر المهراز فاس
ⵜⴰⵖⵡⵍⵉⵏⵜ ⵏ ⵜⵣⵔⴼⵜ ⵎⴰⵔⵔⴰⵣ ⴼⴰⵙ
FACULTE DES SCIENCES DHAR EL MAHRAZ FES

N° d'ordre : 118/2023

Centre des Etudes : ***Sciences et Techniques et Sciences Médicales***

Formation Doctorale : ***STIC***

Discipline : ***Informatique***
Spécialité :***Informatique***
Laboratoire : ***Laboratoire d'Ingénierie, Modélisation et Analyse des Systèmes***

THÈSE DE DOCTORAT

Contribution aux systèmes de recommandation à base de filtrage collaboratif

Présentée par :

El Fazziki Abdellah

Soutenue le 30/12/2023 devant le jury :

Nom et Prénom	Grade	Etablissement	Qualité
Kabbaj Mohammed Nabil	PES	FSDM – USMBA - Fès	Président
Lazaar Mohamed	PES	ENSIAS – UMV - Rabat	Rapporteur& examinateur
Sabri Moulay Abdelouahed	PES	FSDM – USMBA - Fès	Rapporteur& examinateur
Chana Idriss	PH	EST – UMI - Meknès	Rapporteur& examinateur
Guermah Hatim	PH	ENSIAS – UMV - Rabat	Examinateur
Yahyaoui Ali	PES	FSDM – USMBA - Fès	Examinateur
Benbrahim Mohammed	PES	FSDM – USMBA - Fès	Directeur de thèse
El Madani El Alami Yasser	PH	ENSIAS – UMV - Rabat	Co-Directeur de thèse

Année Universitaire : 2022-2023

RESUME

La quantité de données numériques accessibles sur internet est devenue un problème majeur pour les utilisateurs. Heureusement, des systèmes de recommandation ont été créés pour aider les utilisateurs dans diverses activités, telles que l'apprentissage, le shopping, les loisirs, la musique et la lecture, en leur suggérant des produits ou des services personnalisés. Ces systèmes prédisent les préférences des utilisateurs pour des articles qu'ils n'ont pas encore évalués. Actuellement, les systèmes de recommandation jouent un rôle crucial dans les écosystèmes de l'information et du commerce électronique, en permettant aux consommateurs de trier de grandes quantités de données et d'offres de produits. Ils peuvent améliorer l'expérience de l'utilisateur, accroître l'engagement et la fidélité et, en fin de compte, augmenter le chiffre d'affaires des entreprises. Toutefois, ces systèmes sont également confrontés à plusieurs défis, tels que le problème des moutons gris et la rareté des données, qui nécessitent une attention particulière lors de la conception et de la mise en œuvre du système.

Le filtrage collaboratif (CF) est une approche de recommandation populaire qui a fait l'objet de recherches approfondies au cours des deux dernières décennies, ce qui a donné lieu à un ensemble varié d'algorithmes et à une vaste collection d'outils permettant d'évaluer leurs performances. Cette recherche propose une nouvelle approche de recommandation pour traiter les problèmes de mouton gris et de rareté des données, dans le but d'améliorer la précision de la prédiction en déduisant de nouveaux utilisateurs à partir des utilisateurs existants dans les ensembles de données. Cette transformation crée des utilisateurs ayant des préférences opposées à celles des utilisateurs réels, ce qui augmente le nombre d'utilisateurs et résout les deux problèmes mentionnés. Les performances de cette approche

ont été évaluées à l'aide de deux ensembles de données, MovieLens et FilmTrust. Dans l'ensemble, cette recherche contribue au développement de meilleurs systèmes de recommandation capables de relever les défis de la surcharge de données et d'améliorer l'expérience de l'utilisateur.

Mots clés : Système de recommandation, filtrage collaboratif, préférences opposées, problème du mouton gris, problème de parcimonie.

ABSTRACT

The amount of digital data accessible on the Internet has become a major problem for users. Fortunately, recommendation systems have been created to help users in various activities, such as learning, shopping, leisure, music and reading, by suggesting personalised products or services. These systems predict users' preferences for items they have not yet evaluated. Currently, recommender systems play a crucial role in the information and e-commerce ecosystems, allowing consumers to sort through large amounts of data and product offerings. They can improve the user experience, increase engagement and loyalty, and ultimately increase revenue for businesses. However, these systems also face several challenges, such as the grey sheep problem and data scarcity, which require special attention during system design and implementation.

Collaborative filtering (CF) is a popular recommender approach that has been extensively researched over the last two decades, resulting in a diverse set of algorithms and a large collection of tools to evaluate their performance. This research proposes a new recommendation approach to deal with grey sheep and sparsity problems, with the aim of improving prediction accuracy by inferring new users from existing users in the datasets. This transformation creates users with opposite preferences to the real users, which increases the number of users and solves the two problems mentioned. The performance of this approach was evaluated using two datasets, MovieLens and FilmTrust. Overall, this research contributes to the development of better recommender systems that can address the challenges of data overload and improve the user experience.

Key words : Recommendation system, Collaborative filtering, Opposite preferences, Gray sheep problem, Sparsity problem.

DÉDICACES

Ma chère mère et ma chère femme pour leur soutien, affection et amour, leur confiance et patience et pour leurs sacrifices infinis, je le dédie aussi à mes fils ADAM et NIZAR et ma fille INASS.

Mes sœurs et mes frères pour leur encouragement, je leur souhaite une vie pleine de santé et réussite et j'exprime mon amour et mon respect le plus dévoué pour toute ma famille.

Tous mes amis et collègues de travail et particulièrement les plus intimes en témoignage des moments inoubliables, des sentiments purs, et des liens solides qui nous unissent.

Et enfin, tous ceux qui m'ont soutenus et qui me soutiennent encore pour que ce travail soit une déclaration divine de ma profonde estimation et gratitude.

Merci

REMERCIEMENTS

Tout d'abord, je rends grâce au bon DIEU Tout-Puissant pour m'avoir octroyé la force de survivre et la détermination pour surmonter tous les obstacles.

Je tiens à exprimer ma sincère et profonde gratitude envers mes superviseurs, le Professeur BENBRAHIM MOHAMMED et le Professeur EL MADANI EL ALAMY YASSER, pour m'avoir introduit au domaine de la recherche, ainsi que pour leur disponibilité et leur réel intérêt pour ce travail. Leur guidance tout au long de mes études et leurs précieux conseils ont été d'une importance inestimable. Je suis reconnaissant pour leurs encouragements, qui ont été une source inestimable de soutien. Leur lecture attentive de cette thèse et leurs commentaires ont contribué à en améliorer le contenu, et je les en remercie sincèrement.

Je tiens à exprimer mon appréciation et ma gratitude au Professeur MOHAMMED NABIL KABBAJ pour avoir accepté de présider le jury de ma thèse.

J'exprime ma profonde gratitude et mes sincères remerciements au Professeur LAZAAR MOHAMED, au Professeur SABRI MOULAY ABDELOUAHED et au Professeur CHANA IDRISS pour leur temps et leurs efforts dans la révision de ce manuscrit.

Je tiens à remercier le Professeur GUERMAH HATIM et le Professeur YAHYAOUI ALI d'avoir accepté d'examiner cette thèse et pour leurs commentaires utiles et leurs précieuses suggestions.

Afin de n'oublier personne, mes vifs remerciements s'adressent à tous ceux qui m'ont aidée à la réalisation de ce travail

TABLE DES MATIERES

RESUME I

ABSTRACT III

DÉDICACES IV

REMERCIEMENTS V

LISTE DES FIGURES X

LISTE DES TABLEAUX XI

LISTE DES ABREVIATIONS XII

CHAPITRE 1: INTRODUCTION 1

1. Introduction 1
2. Problématique, Hypothèse et motivation 3
   - 2.1. Problématique 3
   - 2.2. Hypothèse 5
   - 2.3. Motivation 6
   - 2.4. Questions de recherche 7
3. Structure de la thèse 7
4. Publications 8
5. Conclusion 9

CHAPITRE 2: ÉTAT DE L'ART 10

1. Introduction 10
2. Histoire des systèmes de recommandation 12
3. Les types des systèmes de recommandation 14
4. Filtrage collaboratif (CF) 16

4.1. Approches basées sur la mémoire 16
4.1.1. Filtrage collaboratif basé sur les utilisateurs (UBCF) 17
4.1.2. Filtrage collaboratif basé sur les items (IBCF) 18
4.1.3. Processus de filtrage collaboratif 19
4.1.3.1. Représentation des données 20
4.1.3.2. Formation de voisinage 20
4.1.3.3. Génération des prédictions 22
4.2. Approches basées sur les modèles 23
4.2.1. Techniques de clustering 24
4.2.2. Arbre de décision 24
4.2.3. Règles d'association 25
4.2.4. Approches de réduction de dimensionnalité 25
4.2.5. Classificateur bayésien 25
4.2.6. Réseaux de neurones 26
4.3. Challenge de filtrage collaboratif 26
4.3.1. Rareté des données ou la matrice creuse 26
4.3.2. Évolutivité 27
4.3.3. Démarrage à froid 27
4.3.4. Mouton gris 28
4.3.5. Attaque de shilling 28
5. Approche basée sur le contenu 29
5.1. Représentation de l'item 30
5.1.1. Modèle d'espace vectoriel basé sur des mots-clés 32
5.1.2. Analyse sémantique basant sur l'ontologie 34
5.1.3. Analyse sémantique basant sur l'encyclopédie 36
5.2. Méthodes d'apprentissage des profils d'utilisateurs 37
5.2.1. Méthodes probabilistes et Naïve bayésienne 38
5.2.2. Pertinence Feedback et algorithme de Rocchio 40
5.2.3. Autre Méthodes 41
6. Approche hybride 43
6.1. Système de recommandation hybride pondéré 43
6.2. Système de recommandation hybride commutation 44
6.3. Système de recommandation hybride Mixte 45

6.4. Système de recommandation hybride Combinaison de caractéristiques 46
6.5. Système de recommandation hybride cascade 47
6.6. Système de recommandation hybride augmentation de caractéristiques 48
6.7. Système de recommandation hybride méta-niveau 49
7. Méthodologies d'évaluation 50
7.1. Evaluations hors ligne (Offline Evaluation) 50
7.1.1. Mesures d'exactitude statistique 51
7.1.1.1. Erreur Absolue Moyenne (MAE) 51
7.1.1.2. Racine d'erreur quadratique moyenne (RMSE) 52
7.1.1.3. Erreur quadratique moyenne (MSE) 53
7.1.1.4. Erreur Absolue Moyenne Normalisée (NMAE) 53
7.1.2. Mesures d'aide à la décision 54
7.2. Evaluation en ligne 57
7.3. Etudes des utilisateurs (user studies) 58
8. Conclusion 58
CHAPITRE 3: FILTRAGE COLLABORATIF BASE SUR L'INFERENCE DES UTILISATEURS IMAGINAIRES 60
1. Introduction 60
2. Problème de moutons gris (gray sheep) 63
3. Notre Approche 65
4. Expérimentation et discussion 69
4.1. Datasets 69
4.2. Expériences 70
4.2.1. Corrélation de Pearson : 70
4.2.2. Cosinus: 71
4.3. Teste statistique 72
5. Conclusion 73
CHAPITRE 4 : FILTRAGE COLLABORATIF AMELIORE A BASE DE MODELE POUR L'ADRESSAGE DU PROBLEME DE MOUTON GRIS 75
1. Introduction 75

2. Travaux connexes 78
3. Notre approche.....83
4. Résultats expérimentaux.....85
4.1. Collecte d'ensembles de données : 86
4.2. Expériences 87
4.3. Résultats et discution.....87
4.4. Inférences statistiques.....95
5. Conclusion 97
CONCLUSION ET TRAVAUX FUTURE.....98
REFERENCES 101

LISTE DES FIGURES

Figure 1: Approches de filtrage collaboratifs :(a) basés sur les utilisateurs ; (b) basés sur les items(Ni et al., 2021) 17
Figure 2 : filtrage collaboratif basé sur l'utilisateur 18
Figure 3: filtrage collaboratif basé sur l'Item 19
Figure 4: Processus de recommandation de filtrage collaboratif 19
Figure 5:image illustrative des collectes des donnés 20
Figure 6: Illustration des systèmes de recommandation basés sur le contenu reposant sur VSM 34
Figure 7: Exemple d'un système de recommandation hybride pondéré 44
Figure 8: Exemple d'un système de recommandation hybride commutatif 45
Figure 9: Illustration des systèmes de recommandation mixte 46
Figure 10: Exemple des systèmes de recommandation hybride combinaison de caractéristiques 47
Figure 11: Exemple des systèmes de recommandation hybride en cascade 48
Figure 12: Exemple des systèmes de recommandation hybride augmentation de caractéristiques 49
Figure 13: Matrice de confusion d'extraction. 55
Figure 14: Exemple de cas de GS dans les techniques basées sur le voisinage 66
Figure 15: Nouveau processus CF basé sur la mémoire 67
Figure 16: Algorithme d'augmentation de la matrice 69
Figure 17: Comparaison de la MAE à l'aide de l'ensemble de données FilmTrust 71
Figure 18: Comparaison de la MAE à l'aide de l'ensemble de données MovieLens 71
Figure 19: Comparaison de la MAE à l'aide de l'ensemble de données FilmTrust 71
Figure 20: Comparaison de la MAE à l'aide de l'ensemble de données MovieLens 71
Figure 21: Nouveau processus de l'approche du CF 84
Figure 22: Dataset divisées 86
Figure 23:Comparaison entre AR et Enriched_AR 90
Figure 24: Comparaison entre SVD et Enriched_SVD 92
Figure 25: Comparaison entre FSVD et Enriched_FSVD 95

LISTE DES TABLEAUX

Tableau 1: Exemple à petite échelle de représentation de données 20

Tableau 2: Exemple à petite échelle de génération de prédictions 23

Tableau 3: Exemple de matrice d'évaluation opposée sur une échelle de 5 points 68

Tableau 4: Comparaison de la valeur p entre la corrélation de Pearson et cosinus par le test de Wilcoxon pour les deux ensembles de données..... 73

Tableau 5: Exemple de matrice d'évaluation opposée sur une échelle de 5 points 85

Tableau 6: Comparaison des RMSE, MSE et MAE entre AR et enriched_AR pour 10 expériences 88

Tableau 7: Comparaison des RMSE, MSE et MAE entre SVD et enriched_SVD pour 10 expériences..... 91

Tableau 8: Comparaison des RMSE, MSE et MAE entre FSVD et enriched_FSVD pour 10 expériences 93

Tableau 9: Comparaison de la valeur p entre RMSE, MSE et MAE par t -test pour les méthodes enrichis et les méthodes traditionnels 96

LISTE DES ABREVIATIONS

GS ***: Mouton gris (gray sheep)***

IA ***: Intelligence Artificiel***

CF ***: Filtrage collaboratif (Collaborative filtering)***

UBCF ***: filtrage collaboratif à base de l'utilisateur (User-Based Collaborative filtering)***

IBCF ***: filtrage collaboratif à base de l'Item (Item-Based Collaborative filtering)***

MAE ***: Erreur Absolue Moyenne (Mean Absolute Error)***

NMAE ***: Erreur Absolue Moyenne Normalisée (Normalized Mean Absolute Error)***

MSE ***: Erreur quadratique moyenne (Mean Squared Error)***

RMSE ***: Erreur quadratique moyenne ((Root Mean Squared Error)***

AR ***: Règles d'association***

SVD ***: Décomposition en valeurs singulières***

FSVD ***: Funk décomposition en valeurs singulières***

LR ***: Régression logistique (LR)***

FM ***: Machines à factorisation (FM)***

SVM ***: Machines à vecteurs de support (SVM)***

ML ***: Apprentissage automatique***

SVM ***: Modèle d'espace vectoriel***

CHAPITRE 1: INTRODUCTION

Ce chapitre sert d'introduction aux systèmes de recommandation, offrant un bref historique de leur évolution. Il expose également notre motivation et la problématique à laquelle nous nous attaquons. En conclusion, il présente la structure détaillée de notre thèse.

1. Introduction

Aujourd'hui, notre monde est fortement axé sur le numérique et l'information, avec une croissance exponentielle de l'utilisation de l'intelligence artificielle (IA), en particulier l'apprentissage automatique et l'apprentissage profond. Les systèmes de recommandation font partie des applications concrètes de l'IA qui ont un impact significatif sur notre vie quotidienne. En raison de la surcharge d'informations sur Internet, il devient de plus en plus difficile pour les utilisateurs de trouver des articles pertinents parmi une multitude d'options, tels que des produits à acheter, des films à regarder, des disques intéressants, des voyages à faire, etc. Les systèmes de recommandation peuvent extraire une liste réduite d'articles correspondant le mieux aux goûts de chaque utilisateur en deux phases : la prédiction et la recommandation proprement dite. Ces systèmes sont devenus indispensables dans de nombreux domaines, tels que la recommandation de sites web, de produits, de films, de musique, de lieux touristiques, de restaurants, ou même pour suggérer de nouveaux profils à recruter ou de nouvelles connexions dans les réseaux sociaux.

Les systèmes de recommandation sont un type de technologie de filtrage de l'information qui fournit des recommandations personnalisées aux utilisateurs en fonction de leurs préférences et de leur comportement. Ces systèmes sont utilisés dans un large

éventail d'applications, du commerce électronique aux médias sociaux en passant par les services de diffusion de musique et de vidéo.

L'objectif d'un système de recommandation est de prédire ce qui pourrait intéresser un utilisateur en fonction de son comportement et de ses préférences antérieures, et de lui présenter des recommandations personnalisées qui répondent à ses besoins et à ses intérêts. Cet objectif est généralement atteint en analysant de vastes ensembles de données sur le comportement des utilisateurs et en utilisant des algorithmes d'apprentissage automatique pour identifier des modèles et des tendances.

Il existe plusieurs types de systèmes de recommandation, notamment le filtrage collaboratif, le filtrage basé sur le contenu et les systèmes hybrides qui combinent les deux approches. Le filtrage collaboratif analyse le comportement et les préférences des utilisateurs afin d'identifier les similitudes entre eux et de recommander des articles que des utilisateurs similaires ont appréciés. Le filtrage basé sur le contenu, quant à lui, se concentre sur les caractéristiques des articles et recommande des articles similaires à ceux qu'un utilisateur a déjà appréciés. Les systèmes hybrides combinent ces approches pour fournir des recommandations plus précises et plus personnalisées.

Les systèmes de recommandation sont devenus de plus en plus importants dans l'ère numérique actuelle, où une quantité écrasante d'informations est disponible et où les utilisateurs cherchent souvent des moyens de filtrer et de hiérarchiser le contenu. En fournissant des recommandations personnalisées, ces systèmes peuvent aider les utilisateurs à découvrir de nouveaux produits et services, à gagner du temps et de l'énergie, et à améliorer leur expérience globale. Dans le même temps, les entreprises et les organisations peuvent bénéficier d'une satisfaction et d'un engagement accrus de la part de leurs clients,

ce qui se traduit par une augmentation des ventes et du chiffre d'affaires.

Cependant, malgré leur importance croissante, les systèmes de recommandation sont confrontés à des défis inhérents, parmi lesquels les problèmes de "sparsity" et de "gray sheep" se démarquent. La "sparsity" se réfère à la rareté des données disponibles pour générer des recommandations précises, tandis que les utilisateurs "gray sheep" présentent des préférences difficiles à catégoriser, posant ainsi des défis supplémentaires pour les systèmes de recommandation traditionnels.

Cette thèse s'engage à explorer en profondeur le domaine des systèmes de recommandation, en mettant particulièrement l'accent sur le filtrage collaboratif. En examinant les aspects techniques, sociaux et éthiques, cette recherche aspire à proposer des solutions innovantes pour surmonter les défis inhérents aux systèmes de recommandation. En visant une compréhension approfondie des problèmes de "sparsity" et des "gray sheep", cette thèse cherche à ouvrir de nouvelles perspectives pour une recommandation plus précise et pertinente, contribuant ainsi à l'évolution continue de cette discipline fondamentale dans le paysage numérique contemporain.

2. Problématique, Hypothèse et motivation

2.1. Problématique

Plusieurs facteurs ont conduit au développement et à l'adoption de systèmes de recommandation parmi ces facteurs on peut cité:

- ✓ La surcharge d'informations : La quantité d'informations disponibles en ligne ayant augmenté de manière exponentielle, La navigation et la recherche de contenus sont devenues de plus en plus difficile pour les utilisateurs. Les systèmes de recommandation permettent de filtrer et de hiérarchiser les

informations, facilitant ainsi la recherche de contenus pertinents et utiles (Resnick & Varian, 1997).

- ✓ Personnalisation : Les utilisateurs attendent des expériences personnalisées qui répondent à leurs besoins et intérêts individuels. Les systèmes de recommandation fournissent des recommandations personnalisées basées sur le comportement antérieur et les préférences de l'utilisateur (Konstan & Riedl, 2012).
- ✓ La concurrence : Sur le marché concurrentiel actuel, les entreprises cherchent des stratégies pour se démarquer de leurs concurrents dans un marché saturé. Les systèmes de recommandation offrent un avantage concurrentiel en proposant des recommandations personnalisées et pertinentes aux utilisateurs (Adomavicius & Tuzhilin, 2005).
- ✓ Big data : La croissance des big data a rendu possible l'analyse de vastes quantités de comportements et de préférences des utilisateurs, ce qui permet des recommandations plus précises et plus personnalisées(Burke, 2002).
- ✓ Satisfaction des clients : En fournissant des recommandations personnalisées, les systèmes de recommandation peuvent améliorer la satisfaction et la fidélité des clients, ce qui se traduit par une augmentation des ventes et du chiffre d'affaires(Ricci et al., 2011).

Ces facteurs ont contribué au développement et à l'adoption des systèmes de recommandation, qui sont devenus un outil essentiel pour les entreprises et les organisations qui cherchent à fournir des recommandations personnalisées et pertinentes à leurs utilisateurs

2.2. Hypothèse

L'hypothèse des systèmes de recommandation est qu'en analysant les comportements passés des utilisateurs, tels que leurs interactions avec les articles, il est possible de faire des prédictions sur les articles qu'ils pourraient être intéressés à consommer à l'avenir. L'hypothèse sous-jacente est que les préférences des utilisateurs sont stables dans le temps et peuvent être déduites de leur comportement, et que des utilisateurs ou des éléments similaires peuvent être identifiés en fonction de modèles de comportement.

Les systèmes de recommandation fonctionnent en exploitant les données pour créer des modèles qui prédisent la probabilité qu'un utilisateur consomme un article. Ces modèles peuvent être basés sur divers types de données, y compris les évaluations des utilisateurs, les commentaires explicites et les commentaires implicites tels que les données de parcours de navigation ou l'historique des achats. Les modèles peuvent utiliser des techniques telles que le filtrage collaboratif, le filtrage basé sur le contenu ou des approches hybrides pour faire des prédictions.

L'hypothèse des systèmes de recommandation est étayée par un grand nombre de recherches qui ont démontré l'efficacité de ces systèmes dans divers domaines, notamment le commerce électronique, le divertissement et les réseaux sociaux. Cependant, il existe également des défis et des limites associés aux systèmes de recommandation.

En résumé, l'hypothèse des systèmes de recommandation est qu'en analysant le comportement des utilisateurs et en identifiant des modèles, il est possible de faire des recommandations personnalisées pertinentes et utiles aux utilisateurs.

2.3. Motivation

Le développement de systèmes de recommandation vise à fournir des recommandations personnalisées aux utilisateurs, en les aidant à découvrir des éléments pertinents tels que des produits, des services, des contenus et des informations, et à s'y intéresser. En exploitant les préférences et le comportement des utilisateurs, ainsi que d'autres données pertinentes, les systèmes de recommandation visent à identifier les éléments qui correspondent aux intérêts et aux besoins des utilisateurs, ce qui permet d'améliorer l'engagement et la satisfaction des utilisateurs.

Voici quelques-unes des principales motivations qui poussent le développement des systèmes de recommandation :

- ✓ Personnalisation : Les systèmes de recommandation visent à fournir des recommandations personnalisées à chaque utilisateur en fonction de ses préférences, de ses intérêts et de ses besoins. Ce faisant, ces systèmes peuvent aider les utilisateurs à découvrir des éléments pertinents de manière plus efficace, ce qui accroît la satisfaction des utilisateurs (Ricci et al., 2015).
- ✓ Augmentation de l'engagement : En fournissant des recommandations personnalisées, les systèmes de recommandation peuvent accroître l'engagement des utilisateurs envers la plateforme ou le service, ce qui se traduit par une fidélité et une rétention accrues (Adomavicius & Tuzhilin, 2005).
- ✓ Amélioration de l'efficacité : Les systèmes de recommandation peuvent aider les utilisateurs à trouver des éléments pertinents plus rapidement et plus facilement, réduisant ainsi le temps et les efforts nécessaires à la recherche manuelle d'éléments

(Burke, 2002).

- ✓ Amélioration de l'expérience utilisateur : En fournissant des recommandations précises et pertinentes, les systèmes de recommandation peuvent améliorer l'expérience globale de l'utilisateur, ce qui se traduit par une satisfaction et une fidélité accrues (R. Chen et al., 2018).
- ✓ Augmentation des revenus : En fournissant des recommandations personnalisées et pertinentes, les systèmes de recommandation peuvent accroître l'engagement des utilisateurs et générer davantage de ventes, ce qui se traduit par une augmentation des recettes pour les entreprises (Chaudhari & Thakkar, 2020).

2.4. Questions de recherche

Cette étude cherche à optimiser le système de recommandation en intégrant de nouveaux concepts dans le filtrage collaboratif. Elle répondra aux questions suivantes

- ✓ **QUESTION 1.** Comment améliorer la qualité des recommandations pour les utilisateurs considérés comme atypiques (gray sheep)?
- ✓ **QUESTION 2.** Comment amplifier la matrice d'évaluation pour résoudre le problème de la matrice creuse afin de garantir des recommandations plus précises et adaptées aux utilisateurs?

3. Structure de la thèse

La structure de la thèse est organisée comme suit :

Chapitre 1 : Nous introduisons les systèmes de recommandation traditionnels et décrivons notre motivation suivie des buts et objectifs. Nous présentons une liste de publications.

Chapitre 2 : Nous proposons une revue détaillée des approches et algorithmes utilisés pour le système de recommandation. Nous présentons également des mesures d'évaluation pour

les systèmes de recommandation.

Chapitre 3 : Nous présentons notre première contribution proposée et nos collections de tests. Nous décrivons les différentes phases de notre approche proposée et expliquons notre configuration expérimentale ainsi les résultats obtenues.

Chapitre 4 : Nous détaillons notre deuxième contribution, y compris la collection des données de tests utilisées. Nous procédons également à une comparaison approfondie de nos résultats avec certaines techniques de filtrage collaboratif traditionnelles.

Chapitre 5 : Nous présentons la conclusion générale et notre contribution ainsi que les travaux futurs discutés.(Maidin et al., 2024)

4. Publications

- El Fazziki, A., EL Madani El alami, Y., El hassouni, J. & Benbrahim, M. "ENHANCING COLLABORATIVE FILTERING : ADDRESSING SPARSITY AND GRAY SHEEP WITH OPPOSITE USER INFERENCE" **102**, (2024). https://www.jatit.org/volumes/Vol102No3/9Vol102No3.pdf
- El Fazziki, Y. E. M. El Alami, J. Elhassouni, O. El Aissaoui, and M. Benbrahim, "Employing opposite ratings users in a new approach to collaborative filtering," Indones. J. Electr. Eng. Comput. Sci., vol. 25, no. 1, pp. 450–459, 2022.
- El Fazziki, A. El, Aissaoui, O. El, Alami, Y. E. M. El, Allioui, Y. El, & Benbrahim, M. (2019). A new collaborative approach to solve the gray-sheep users problem in recommender systems. 2019 3rd International Conference on Intelligent Computing in Data Sciences, ICDS 2019, 1–4. https://doi.org/10.1109/ICDS47004.2019.8942256.
- El Fazziki A., El Aissaoui O., El Madani El Alami Y., Benbrahim M., El Allioui Y.

(2020) A Novel Collaborative Filtering Approach Based on the Opposite Neighbors' Preferences. In: Bhateja V., Satapathy S., Satori H. (eds) Embedded Systems and Artificial Intelligence. Advances in Intelligent Systems and Computing, vol 1076. Springer, Singapore. https://doi.org/10.1007/978-981-15-0947-6_77

- El Fazziki, A., El Madani El Alami, Y., El Aissaoui, O., El Allioui, Y., & Benbrahim, M. (2020). Improving Collaborative Filtering Approach by Leveraging Opposite Users. In Advances in Intelligent Systems and Computing. https://doi.org/10.1007/978-3-030-36653-7_14.

5. Conclusion

En conclusion de ce chapitre introductif, nous avons identifié la problématique cruciale liée à la surabondance d'informations, exposant la nécessité de systèmes de recommandation efficaces. La motivation sous-jacente à cette recherche réside dans la conviction que ces systèmes jouent un rôle central dans l'amélioration de l'expérience utilisateur. Nos questions de recherche détaillent le chemin que prendra notre exploration, tandis que la structure de la thèse offre un cadre organisé pour cette quête intellectuelle. Enfin, la perspective de futures publications souligne notre engagement envers le partage et la contribution au corpus de connaissances dans ce domaine crucial des systèmes de recommandation.

Le chapitre suivant proposera une revue de la littérature détaillée sur les systèmes de recommandation, offrant une exploration approfondie des divers types qui existent.

CHAPITRE 2: ÉTAT DE L'ART

Ce chapitre offre une plongée approfondie dans l'état actuel de l'art des systèmes de recommandation. En explorant les avancées récentes, les tendances émergentes et les défis persistants, nous établirons un panorama détaillé du paysage actuel des systèmes de recommandation. Cette revue exhaustive vise à positionner notre recherche dans le contexte global, mettant en lumière les lacunes existantes et les opportunités prometteuses qui guideront notre exploration ultérieure.

1. Introduction

Les systèmes de recommandation sont des outils intelligents qui peuvent recommander des choses aux utilisateurs en fonction de leurs préférences (Polatidis & Georgiadis, 2017), en particulier personnaliser le système de recommandation. Le premier but des systèmes de recommandation est de faciliter la tâche des utilisateurs en leur fournissant des items répondant à leurs besoins. La clé des systèmes de recommandation personnalisés réside dans les préférences de l'utilisateur. En fait, les préférences permettent aux systèmes de recommandation de comprendre les besoins et le comportement des utilisateurs. Ils sont principalement basés sur l'historique des utilisateurs (notes, clics, etc.), ces interactions sont divisées en deux catégories : les données implicites (par exemple, les clics, les achats) et les données explicites (par exemple, les évaluations, les votes). Les données de retour d'informations explicites sont plus largement utilisées dans les domaines de recherche du système de recommandation (Adomavicius & Tuzhilin, 2005)(Takács & Tikk, 2012).

Avec le problème croissant de la surcharge d'informations, les systèmes de recommandation sont devenus indispensables pour que les utilisateurs trouvent facilement des informations, des produits ou des services précis qu'ils recherchent. Ces systèmes ont pour rôle de filtrer les informations entrantes en transmettant les flux pertinents à l'utilisateur et en bloquant ceux qui ne le sont pas (Ortega et al., 2018).

Au cours des dernières décennies, les industries en ligne ont intensivement utilisé les systèmes de recommandations pour revendiquer leur place sur le marché et améliorer leur gestion de la relation client. Par exemple, les systèmes de commerce électronique tels qu'Amazon (Linden et al., 2017), les systèmes de voyage tels que TravelJoy (Samah et al., 2019), les plateformes de streaming de films tels que Netflix (Gomez-uribe & Hunt, 2015) et les applications musicales (Kasinathan et al., 2019)(Celma, 2008)(Callan et al., 2003) ont connu un grand succès en rendant les divertissements et les achats facilement accessibles et en offrant une expérience incroyable aux utilisateurs, en particulier pendant la pandémie de Covid-19.

De nombreuses approches de système de recommandation ont été proposées et développées afin de répondre aux besoins croissants des utilisateurs et de surmonter les problèmes rencontrés dans le processus de recommandation.

Nous allons présenter, dans le cadre de ce chapitre, la définition de la problématique de recommandation telle qu'elle a été proposée dans la littérature. Par la suite, nous allons présenter les trois types d'approches qui ont orienté l'ensemble des recherches effectuées. Ainsi que les challenges de filtrage collaboratif et à la fin nous citons les méthodologies d'évaluation des approches des systèmes de recommandation.

2. Histoire des systèmes de recommandation

En 1992, Belkin et Croft (Belkin & Croft, 1992) ont comparé le filtrage et la recherche d'informations, cette dernière étant la technologie utilisée dans les moteurs de recherche et la première étant la base des systèmes de recommandation. À peu près à la même époque, Goldberg et al (D. Goldberg et al., 1992) ont créé le système Tapestry, qui était le premier système de filtrage d'informations utilisant le filtrage collaboratif avec évaluation humaine. Sur la base de ces travaux, des chercheurs du MIT et de l'UMN ont développé GroupLens (Paul Resnick, Neophytos Lacovou, Mitesh Suchak, Peter Bergstrom, 1999), un service de recommandation de nouvelles qui utilise un modèle de filtrage collaboratif basé sur l'utilisateur. Le laboratoire de recherche GroupLens, fondé par le professeur John Riedl à l'UMN, est devenu un leader dans le domaine des études sur les systèmes de recommandation. Des technologies de recommandation similaires ont été développées pour la musique et la vidéo avec Ringo (Shardanand & Maes, 1995) et Video Recommender (Hill et al., 1995), respectivement. Avec l'émergence du commerce électronique, l'industrie a reconnu la valeur des systèmes de recommandation, la première entreprise à proposer un moteur de recommandation marketing étant Net Perceptions (Riedl, n.d.) en 1996. Schafer et al (J. Ben Schafer, Joseph Konstan, 1999) ont expliqué comment les systèmes de recommandation contribuent à augmenter les ventes des sites de commerce électronique en analysant six sites web. Les études universitaires et les applications industrielles des systèmes de recommandation ont progressé de concert depuis lors. En 1997, le laboratoire de recherche GroupLens a lancé le projet MovieLens (Harper & Konstan, 2015), qui a entraîné la première version du modèle de recommandation avec l'ensemble de données EachMovie. Les jeux de données MovieLens suivants ont été publiés de 1998 à 2019 et

sont devenus un jeu de données populaire pour les études de recommandation. Les technologies de filtrage collaboratif ont dominé les applications et les études sur les systèmes de recommandation avant 2005, notamment le filtrage collaboratif basé sur l'utilisateur (Breese et al., 1998) (Jonathan L. Herlocker, Joseph A. Konstan, Al Borchers, 1999), le filtrage collaboratif basé sur l'item (Linden et al., 2003) (Sarwar et al., 2001) et le filtrage collaboratif basé sur la décomposition en valeurs singulières (SVD) (Sarwar et al., 2000). Après le prix Netflix, de 2006 à 2009, les modèles de factorisation matricielle ont fait l'objet d'études approfondies (Koren, 2008) (Koren, 2010). Au cours de cette période, certains chercheurs ont plaidé en faveur d'un changement dans l'évaluation des systèmes de recommandation, passant des mesures de précision conventionnelles à une évaluation centrée sur l'utilisateur (McNee et al., 2006).

En 2007, en raison de l'essor rapide de la recherche fondamentale et des applications commerciales dans le domaine des systèmes de recommandation, l'Université du Minnesota (UMN) a accueilli la première conférence ACM sur les systèmes de recommandation. Aujourd'hui, l'ACM RecSys est l'une des plus importantes conférences universitaires annuelles consacrées à l'étude des systèmes de recommandation. En 2007, Richardson et al. ont proposé un modèle de régression logistique (LR) qui a permis de réduire de 30 % l'erreur d'estimation du taux de clics. Depuis lors, les modèles LR ont été continuellement améliorés sous divers aspects, y compris les méthodes d'optimisation et l'ingénierie automatique des caractéristiques. En 2010, Rendle a proposé les machines à factorisation (FM) qui combinent les avantages des machines à vecteurs de support (SVM) et des modèles de factorisation. Basées sur les FM, les machines de factorisation conscientes des champs (FFM) ont été proposées par Juan et al. qui prennent en compte les champs des

caractéristiques lors de la modélisation du poids de chaque paire de caractéristiques. Ces dernières années, les réseaux neuronaux profonds ont émergé à la fois dans les universités et dans l'industrie pour les modèles de recommandation.

Pour résoudre le problème de la reproductibilité dans les études sur les modèles de recommandation, des critères de référence ouverts ont été mis au point, tels que FuxiCTR pour la prédiction du CTR et des mesures d'évaluation reproductibles et équitables pour les algorithmes de recommandation top-N basés sur le feed-back implicite. Ces dernières années, un nombre croissant d'études ont été consacrées à la recommandation inspirée par l'inférence causale afin de remédier aux biais des systèmes de recommandation. Schnabel et al. ont proposé une approche du traitement des biais de sélection en adaptant les modèles et les estimations de l'inférence causale.

Le présent document ne vise pas à fournir un examen complet de tous les aspects des systèmes de recommandation, étant donné qu'il existe des études approfondies sur des sujets de recherche connexes. Il se concentre plutôt sur deux aspects des systèmes de recommandation fondés sur les recherches et les expériences industrielles des auteurs : les modèles de recommandation pratiques et les architectures des systèmes de recommandation typiques. Enfin, les auteurs évoquent brièvement quelques idées sur l'avenir des systèmes de recommandation.

3. Les types des systèmes de recommandation

La recommandation est une sous-classe de système de filtrage d'informations qui cherche à prédire la « note » ou la « préférence » qu'un utilisateur donnerait à un item. Le système de recommandation est devenu plus populaire ces dernières années.

Un système de recommandation tente d'estimer/prédire une fonction de notation R sur la base d'un ensemble initial de notations. La fonction de notation R peut être spécifiée à l'aide de l'équation 1 suivante.

$$R = User \times Item \rightarrow Rating \quad (1)$$

Le problème de base du système de recommandation est d'estimer ou de prédire une fonction d'utilité qui peut aider à prédire comment un utilisateur aimera un item. Dans la fonction utilitaire, U est représenté comme un ensemble d'utilisateurs U := {users}, I est représenté comme un ensemble d'Items recommandables I := {recommendableitems} où F:= fonction utilitaire

$$F = U \times I \rightarrow R \quad (2)$$

où R := {Items recommandés}, et pour chaque utilisateur u, nous voulons choisir les items i qui maximisent f , comme le montre l'équation 3 suivante.

$$u \in U \quad i´u = argmaxf\ F(u, i) \quad (3)$$

Les systèmes de recommandations s'articulent sur deux composants de base : les utilisateurs et les items. Dans ce qui suit, nous allons adopter le formalisme suivant : l'espace des utilisateurs sera noté $U = \{u_1 \dots u_n\}$ et l'espace des items sera noté $I = \{i_1 \dots i_p\}$.

Selon (Adomavicius & Tuzhilin, 2005), trois grands types de systèmes de recommandation ont été proposés dans la littérature:

1. Filtrage Collaboratif (CF) (El Madani El Alami et al., 2016),
2. Approche basés sur le contenu (M. J. Pazzani & Billsus, 2007),
3. Approche hybrides (Burke, 2002)(El Madani El Alami et al., 2015).

4. Filtrage collaboratif (CF)

L'approche Filtrage collaboratif (CF) est l'approche la plus utilisée dans les systèmes de recommandation en raison de leur efficacité et de leur simplicité (Ekstrand, 2011). Dans la littérature, les approches de CF sont basées sur une simple intuition. Ils supposent que de bonnes recommandations peuvent être dérivées d'utilisateurs partageant les mêmes intérêts et préférences. Ces préférences peuvent être exprimées de plusieurs manières, soit en utilisant des évaluations basées sur les intérêts des utilisateurs (Jawaheer et al., 2006), soit en déduisant du comportement des utilisateurs, en suivant leur historique sur le web (l'achat et le temps passé sur les pages web), ce que l'on appelle le retour implicite (Y. Hu et al., 2008). Ils peuvent être exprimés sous la forme d'une matrice appelée matrice de notation (Hernando et al., 2016). C'est la base pour créer des modèles de prédiction efficaces et des profils d'utilisateurs (Isinkaye et al., 2015a). Dans la littérature, il existe deux approches principales de CF : basée sur la mémoire et basée sur un modèle.

4.1. Approches basées sur la mémoire

Les approches de recommandation CF basées sur la mémoire exploitent directement les préférences des utilisateurs (Gras et al., 2016) (établissant des relations similaires entre les utilisateurs ou les items sur la base de la matrice de notation utilisateur-item). Les techniques révélées dans l'approche basée sur la mémoire sont considérées comme le premier algorithme de CF (Breese et al., 1998). Les recommandations s'appliquent facilement à la matrice de notation. Comme illustrer dans la figure 1, les techniques utilisées dans le CF basé sur la mémoire permettent d'obtenir la similarité pour calculer la distance entre deux utilisateurs (approche basée sur l'utilisateur), ou entre deux items (approche basée sur l'item), selon les évaluations, dans la matrice de notation (Sarwar B, Karypis G,

Konstan J, 2001).

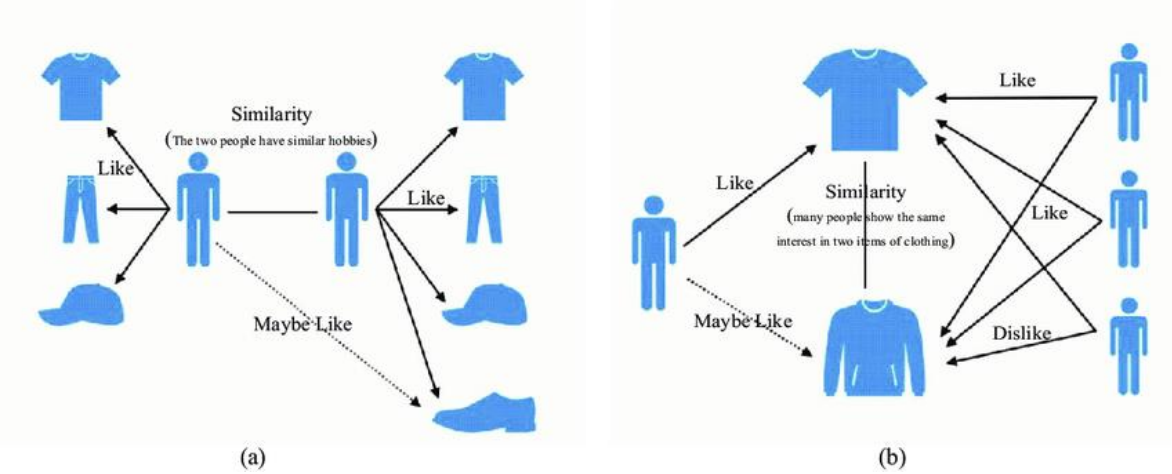


Figure 1: Approches de filtrage collaboratifs :(a) basés sur les utilisateurs ; (b) basés sur les items(Ni et al., 2021)

4.1.1. Filtrage collaboratif basé sur les utilisateurs (UBCF)

L'approche de filtrage collaboratif basée sur l'utilisateur consiste à prédire à l'utilisateur cible des items qui sont déjà des items d'intérêt pour d'autres utilisateurs similaires à l'utilisateur cible. (Walunj & Sadafale, 2013), dans la figure 2 l'utilisateur 1 et l'utilisateur 3 ont un comportement de préférence très similaire. Si l'utilisateur 1 aime l'item A, l'UBCF peut recommander l'item A à l'utilisateur 3. L'UBCF a besoin des scores de notation explicites des items évalués par les utilisateurs (Sarwar B, Karypis G, Konstan J, 2001) pour calculer les similitudes entre les utilisateurs et exploite les algorithmes des k-plus proches voisins pour trouver les voisins les plus proches en fonction de similitudes de l'utilisateur. Et puis, il génère une prédiction en termes d'items en combinant les scores de notation de l'utilisateur voisin sur la base d'une moyenne pondérée par similarité(S. Wei et al., 2012).

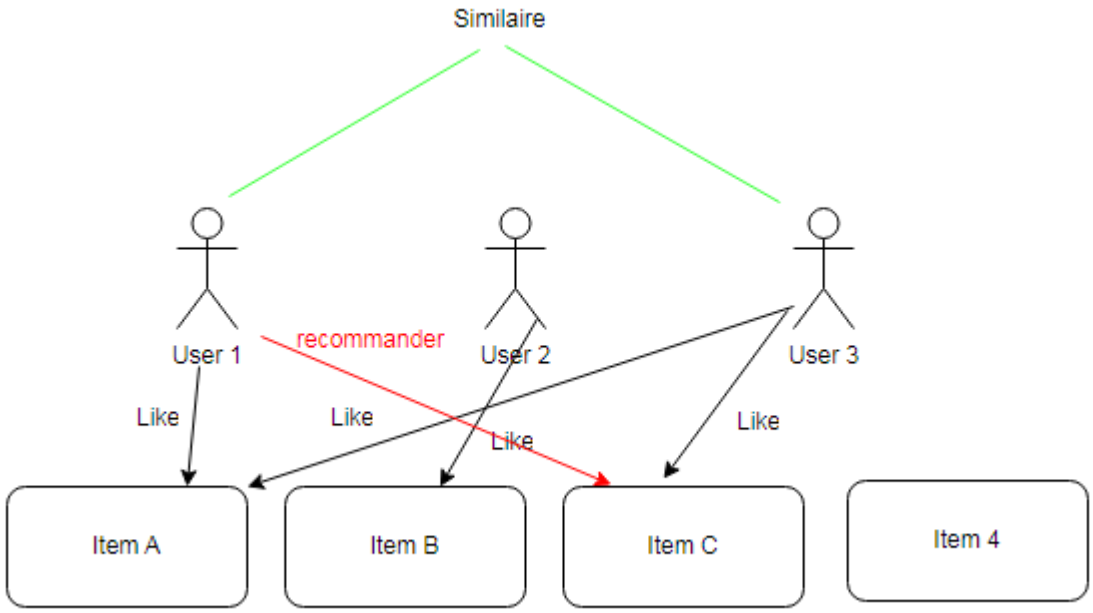


Figure 2 : filtrage collaboratif basé sur l'utilisateur

4.1.2. Filtrage collaboratif basé sur les items (IBCF)

L'approche de filtrage collaboratif basée sur les items consiste à prédire les items en recherchant les similitudes entre les items et d'autres items déjà associés à l'utilisateur. (Walunj & Sadafale, 2013), par exemple dans la figure 3 disons que l'item A et l'item C sont très similaires. Si un utilisateur aime l'article A, l'IBCF peut recommander l'article C à l'utilisateur. L'IBCF à besoin d'un ensemble d'items que l'utilisateur cible a déjà évalué pour calculer les similitudes entre les items et un item cible. Et puis, il génère une prédiction en termes d'item cible en combinant les préférences précédentes de l'utilisateur cible en fonction de ces similitudes d'items (S. Wei et al., 2012). Dans IBCF, les données de préférences des utilisateurs peuvent être collectées de deux manières. La première est que l'utilisateur attribue explicitement une note à l'item dans une certaine échelle numérique. L'autre est qu'il analyse implicitement les enregistrements d'achat ou le taux de clics de

l'utilisateur (Sarwar B, Karypis G, Konstan J, 2001).

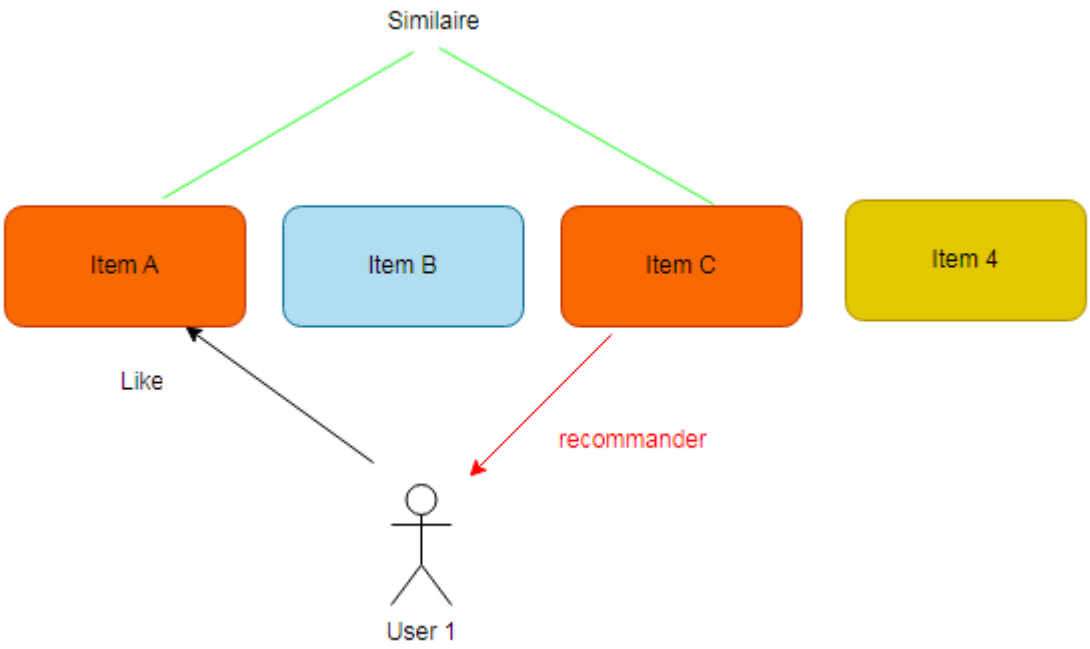


Figure 3: filtrage collaboratif basé sur l'Item

4.1.3. Processus de filtrage collaboratif

L'approche basée sur la mémoire est basée sur trois étapes présentées dans la figure 13 ci-dessous (Bhaidani, 2008):



Figure 4: Processus de recommandation de filtrage collaboratif

La première étape avant d'entrer dans le processus de recommandation de filtrage collaboratif consiste à collecter les données des utilisateurs ayant besoin de recommandations. Ces données (les films évalués dans ce cas) servent de requête à l'algorithme.



Figure 5:image illustrative des collectes des donnés

4.1.3.1. Représentation des données

La deuxième étape de CF consiste à construire la matrice d'évaluation et à remplir les valeurs vides. En fait, dans la plupart des cas, la matrice de notation est généralement remplie car les utilisateurs ne notent pas régulièrement les items (Paterek, 2007). La technique la plus utilisée dans le FC consiste à remplacer les cases vides de la matrice par les notes moyennes des utilisateurs.

Tableau 1: Exemple à petite échelle de représentation de données

	I_1	I_2	I_3	I_4
U_1	5		3	
U_2		4	2	5
U_3		5	2	

4.1.3.2. Formation de voisinage

Dans cette étape, nous recherchons le voisinage des utilisateurs les plus similaires à l'aide d'une métrique de similarité. Il existe différentes mesures pour obtenir la similitude. Cependant, les plus étendus sont le coefficient de corrélation de Pearson et la similarité en cosinus (Breese et al., 1998).

4.1.3.2.1.Coefficient de corrélation de Pearson

Le coefficient de corrélation de Pearson entre les item i et j est calculée comme suit :

$$sim_{a,b} = \frac{\sum_{j=1}^{n}(r_{aj} - \bar{r_a})(r_{bj} - \bar{r_b})}{\sqrt{\sum_{j=1}^{n}(r_{aj} - \bar{r_a})^2 \sum_{j=1}^{n}(r_{bj} - \bar{r_b})^2}} \quad (4)$$

Cette mesure de similarité est souvent utilisée dans la littérature (Adomavicius & Tuzhilin, 2015) (Rao & Talwar, 2008). Cependant, il y a un problème dans le cas où deux items n'ont qu'un seul utilisateur en commun. Supposons que cet utilisateur ait attribué la même note aux deux items, alors la similarité sera égale à 1 (la valeur maximale). De plus, il en sera de même si un millier d'utilisateurs attribuent les mêmes notes à i et j.

4.1.3.2.2.Cosinus

La mesure de similarité du cosinus est une technique utilisée pour évaluer la similitude entre deux vecteurs dans un espace vectoriel. Elle mesure l'angle entre ces vecteurs et produit un score compris entre -1 (complètement opposés) et 1 (entièrement similaires), avec 0 indiquant une indépendance linéaire. Cette mesure est fréquemment employée dans des domaines tels que le traitement du langage naturel, la recherche d'information et le filtrage collaboratif.

La mesure du cosinus est définie comme suit :

$$sim_{a,b} = \cos(a,b) = \frac{a\ b}{\|a\|\|b\|} = \frac{\sum_{j=1}^{n} r_{aj} * r_{bj}}{\sqrt{\sum_{j=1}^{n} {r_{aj}}^2} * \sqrt{\sum_{j=1}^{n} {r_{bj}}^2}} \quad (5)$$

4.1.3.2.3.Jacard

La similarité de Jacard entre les deux ensembles A et B est |A∩B| / |AUB|, c'est-à-dire le rapport de la taille de l'intersection de A et B sur la taille de leur union. Il peut être noté

par la formule suivante :

$$sim_{J(A,B)} = \frac{|A \cap B|}{|A \cup B|} \quad (6)$$

4.1.3.2.4.Distance euclidienne

La méthode de distance euclidienne est basée sur la distance entre les items. Il forme des coordonnées pour mettre des valeurs de préférence entre les items et mesure la distance euclidienne entre chaque point. Lorsque la valeur de la distance entre deux points i et j, (i, j), est grand, cela signifie que les deux points ne sont pas similaires. Lorsque (i, j) est petit, cela signifie que deux points sont similaires. Ceci est la formule de distance euclidienne est donnée ci-dessous :

$$sim(i,j) = \sqrt{\sum_{k=1}^{n}\left(R_{k,i} - R_{k,j}\right)^2} \quad (7)$$

$R_{k,\text{i}}$ est la note de l'item cible i donnée par l'utilisateur k. $R_{k,\text{j}}$ est la note de l'autre item j donnée par l'utilisateur k. n est le nombre total d'évaluation d'utilisateurs pour l'item i et l'item j.

4.1.3.3. Génération des prédictions

Dans cette étape, après avoir sélectionné les k-voisins les plus proches pour l'utilisateur actif en utilisant le degré de similitude basé sur la matrice d'évaluation, le processus CF génère des prédictions pour les items invisibles. La prédiction peut être générée à l'aide de la formule suivante :

$$p_{s,i} = \bar{r}_s + \frac{\sum_{p=1}^{k}(r_{p,i} - \overline{r_p}) * sim_{s,p}}{\sum_{p=1}^{k}|sim_{s,p}|} \quad (8)$$

Dans cette formule, nous calculons les prédictions de notation pour tous les items qui n'ont pas encore été vus par les utilisateurs. Nous utilisons la technique KNN où K représente le nombre de voisins les plus proches et $\bar{r}_s$ représente la note moyenne s de l'utilisateur.

Tableau 2: Exemple à petite échelle de génération de prédictions

User-Item Rating

	I_1	I_2	I_3	I_4
U_1	5	3	3	4
U_2	3	4	2	5
U_3	4	5	2	5

4.2. Approches basées sur les modèles

Dans le filtrage collaboratif basé sur des modèles, des algorithmes d'apprentissage automatique sont utilisés pour développer des modèles et trouver des modèles basés sur des données d'entraînement. Les algorithmes d'apprentissage automatique sont généralement basés sur la notion de recherche de tendances, de modèles ou d'anomalies dans un ensemble de données, et pourraient être avantageusement utilisés pour analyser des ensembles de données volumineux et complexes (Smyth, 2007). Cette technologie pourrait être utilisée pour regrouper certains points de données, trouver des anomalies dans les données ou pour prédire un attribut parmi plusieurs paramètres connus, en fonction des tendances et des modèles trouvés dans les données d'apprentissage (Salakhutdinov et al., 2007).

Dans les méthodes basées sur des modèles, les algorithmes dépendent de la phase d'apprentissage, contrairement à la CF basée sur la mémoire. Cette technique est couramment utilisée pour prédire les données réelles à l'aide de nombreux algorithmes CF basés sur des modèles. Ces méthodes sont fortement inspirées des techniques d'apprentissage automatique telles que les techniques de clustering (Tsai & Hung, 2012), les approches de réduction de dimensionnalité (Paterek, 2007), les machines à vecteurs de support et les réseaux de neurones (Agrawal & Agrawal, 2015).

4.2.1. Techniques de clustering

Le clustering est une méthode qui est utilisée pour partitionner un ensemble de données en no. de grappes. Une bonne méthode de clustering signifie une similarité intra-cluster élevée et une similarité inter-clusters faible. Dans le système de recommandation, les utilisateurs peuvent participer partiellement à différents clusters et le degré de participation peut être calculé en prenant la moyenne des clusters (Sahoo et al., 2019).

4.2.2. Arbre de décision

L'arbre de décision crée un graphique semblable à une structure arborescente qui est construit en considérant l'ensemble de données d'apprentissage dans lequel les étiquettes de classe sont connues. Cet arbre peut être utilisé pour classer les données de test. L'arbre de décision est un type de classificateur qui gère et classe les exemples précédents non vus (B. K. Mishra et al., 2018).

4.2.3. Règles d'association

Exploration de règles d'association : les algorithmes d'exploration de règles d'association génèrent des règles d'association qui décident de la relation entre les items d'une transaction. La règle d'association A->B signifie que l'ensemble d'items A prédit l'ensemble d'items B. Cette règle peut être adaptée à un modèle de recommandation pour prédire l'utilisateur et l'item(Jooa et al., 2016).

4.2.4. Approches de réduction de dimensionnalité

SVD est une méthode bien connue pour établir des facteurs latents dans le cadre des systèmes de recommandation pour travailler sur les problèmes rencontrés par la technique de filtrage collaboratif. Ces techniques sont devenues populaires en raison d'une bonne évolutivité et d'une meilleure précision prédictive. Il s'agit d'une technique de filtrage de recommandation populaire qui décompose la matrice **A** en trois matrices sous la forme $\mathbf{A} = \mathbf{USV^T}$ où **U** et **V** sont deux matrices orthogonales de taille m × r et m × y, respectivement ; y est le rang de la matrice **A**. **S** est une matrice diagonale de taille **r** × **r**. Les éléments diagonaux contiennent toutes les valeurs singulières de la matrice **A**. Il décompose ensuite la matrice **S** pour obtenir une matrice $\mathbf{S_k}$, $\mathbf{k} < \mathbf{r}$ montrant les **k** éléments les mieux notés (plus grandes valeurs diagonales) (Sahoo et al., 2019).

4.2.5. Classificateur bayésien

Le classificateur bayésien est utilisé pour résoudre un problème de classification basé sur la probabilité conditionnelle et le théorème de Bayes. Le classificateur bayésien est utilisé pour prédire la classe en considérant la probabilité de la classe par rapport à un attribut particulier en appliquant le théorème de Bayes. Ce classificateur

est généralement utile lorsque les préférences des utilisateurs changent en ce qui concerne le temps nécessaire à la construction du modèle.

4.2.6. Réseaux de neurones

Réseau de neurones est un réseau de nombreux neurones connectés disposés en différentes couches. Les facteurs de poids et de biais sont associés à chaque neurone de chaque couche. Chaque neurone a une fonction de transfert à travers laquelle il mesure l'entrée, traite cette entrée et donne la sortie.

4.3. Challenge de filtrage collaboratif

Le filtrage collaboratif, qui repose sur la recommandation d'éléments en fonction des comportements antérieurs des utilisateurs, est une approche puissante pour la personnalisation des recommandations. Cependant, il est confronté à plusieurs défis qui limitent parfois son efficacité. Nous citons les principaux challenges de filtrage collaboratif.

4.3.1. Rareté des données ou la matrice creuse

La rareté des données ou la matrice creuse est avant tout un défi informatique. Prenons l'exemple d'un système de recommandation qui fait la promotion de milliers de produits auprès de centaines, voire de milliers de clients. Si les données relatives à l'interaction entre l'utilisateur et le produit étaient enregistrées dans une matrice, il s'agirait d'une quantité massive de données comprenant de nombreux zéros (la plupart des utilisateurs ne s'intéressent qu'à un sous-ensemble sélectionné de produits). Un format de matrice clairsemé pourrait être une meilleure façon de stocker ces données (qui n'enregistre que les non-zéros). Cela pourrait faciliter le

stockage, mais il faudrait quand même un algorithme capable de traiter des données représentées de manière aussi éparse, ce qui n'est pas toujours disponible dans le commerce (K. Yu et al., 2004).

4.3.2. Évolutivité

Même si les utilisateurs sont très actifs dans une base de données d'évaluation des utilisateurs et des articles, il n'y a que quelques évaluations par rapport au nombre total d'articles accessibles. En général, l'ensemble de la base de données d'évaluation fait l'objet d'une recherche dans le cadre du filtrage collaboratif, ce qui entraîne une faible évolutivité au fur et à mesure que des personnes et des objets sont ajoutés à la base de données (Pelikan et al., 2002). Lors du traitement ou de l'analyse de ces données à grande échelle, les systèmes traditionnels de services de recommandation souffrent souvent de problèmes d'évolutivité et d'efficacité. Pour contourner ces problèmes, un système de recommandation unique basé sur l'algorithme de filtrage collaboratif est construit.

4.3.3. Démarrage à froid

Le démarrage à froid est un problème dans les systèmes d'information informatisés qui incluent une certaine modélisation automatisée des données. Il concerne spécifiquement le problème du système qui est incapable de tirer des conclusions sur des personnes ou des produits pour lesquels il n'a pas encore acquis d'informations adéquates (M. Zhang et al., 2014). Le problème du démarrage à froid, qui explique les difficultés à donner des suggestions lorsque les utilisateurs ou les objets sont nouveaux, continue d'être un obstacle important pour la FC.

Traditionnellement, ce défi a été relevé en effectuant une procédure d'entretien supplémentaire pour déterminer le profil de l'utilisateur (de l'objet) avant de faire des suggestions (Volkovs et al., 2017).

4.3.4. Mouton gris

Les consommateurs qui ne sont pas systématiquement en accord ou en désaccord avec un groupe de personnes sont appelés moutons gris, et ils ne profitent pas de la FC. La catégorie opposée est celle des moutons noirs, dont les intérêts inhabituels rendent les suggestions difficiles. Bien qu'il s'agisse d'une défaillance du système de recommandation, les recommandeurs non électroniques ont également des problèmes significatifs dans ces cas, le mouton noir est donc une défaillance appropriée. (Claypool et al., 1999) ont présenté une approche hybride qui combine des recommandations basées sur le contenu et des recommandations CF en basant une prédiction sur une moyenne pondérée des prédictions basées sur le contenu et des prédictions CF.

Dans cette approche, les pondérations des prédictions basées sur le contenu et sur les FC sont calculées par utilisateur, ce qui amène le système à évaluer l'équilibre optimal des recommandations basées sur le contenu et des FC pour chaque utilisateur, contribuant ainsi à la résolution du problème des moutons gris (Ghazanfar & Prügel-Bennett, 2014).

4.3.5. Attaque de shilling

Les gens peuvent faire beaucoup de suggestions positives pour leurs articles et beaucoup de mauvaises recommandations pour leurs rivaux dans des circonstances

où n'importe qui peut soumettre des recommandations. Les systèmes de CF doivent inclure des sauvegardes pour éviter ce type de comportement (Canny, 2002). Des modèles d'attaques de shilling pour les systèmes de CF ont été reconnus et leur efficacité a été étudiée. Ils ont découvert que l'algorithme de CF basé sur les articles était substantiellement moins affecté par les attaques que l'algorithme de CF basé sur l'utilisateur, et ils proposent que de nouvelles méthodes pour évaluer et détecter les attaques de shilling sur les systèmes de recommandation soient développées (Mobasher et al., 2005). (Gunes et al., 2013) ont étudié les modèles d'attaque pour l'escroquerie des systèmes de CF basés sur les items, et d'autres systèmes de CF tels que les systèmes de CF hybrides et les systèmes de CF basés sur des modèles ont été considérés comme capables de fournir des réponses partielles au problème d'injection de biais. (Burke & Hurley, 2011) ont contribué à la résolution du problème de l'injection de biais en examinant la robustesse, c'est-à-dire la résistance d'un système de recommandation à des changements potentiellement nuisibles de la matrice de notation des utilisateurs.

5. Approche basée sur le contenu

Le filtrage basé sur le contenu repose sur les techniques qui utilisent les caractéristiques descriptives d'un item pour fournir des informations sur les préférences des utilisateurs. Les items les plus similaires à ceux préférés précédemment par l'utilisateur sont les éventuels items recommandés. Avec ces techniques le profil d'utilisateur se construit à base des points en communs déduits des items évalués positivement par l'utilisateur (Balabanovi & Shoham, 1997). Les systèmes de recommandation basés sur le contenu se situe à l'intersection de nombreux sujets informatiques, en particulier l'intelligence

artificielle et la recherche d'informations (Baeza-Yates & Ribeiro-Neto, 1999).

Le besoin d'informations de l'utilisateur est représenté par son propre profil. Les items à recommander peuvent être très différents selon le nombre et les types d'attributs utilisés pour les décrire. Chaque item peut être décrit par le même petit nombre d'attributs avec un ensemble de valeurs connu, mais cela n'est pas approprié pour des items, tels que des pages Web, des actualités, des e-mails ou des documents, décrits par un texte non structuré. Dans ce cas, il n'y a pas d'attributs avec des valeurs bien définies, et l'utilisation de techniques de modélisation de documents avec des racines dans la recherche IR est souhaitable.

Du point de vue de l'intelligence artificielle, la tâche de recommandation peut être présentée comme un problème d'apprentissage qui exploite les connaissances passées sur les utilisateurs. Dans leur forme la plus simple, les profils d'utilisateurs se présentent sous la forme de mots-clés ou de règles spécifiés par l'utilisateur et reflètent les intérêts à long terme de l'utilisateur. Souvent, il est conseillé au recommandeur d'apprendre le profil de l'utilisateur plutôt que d'imposer à l'utilisateur d'en fournir un. Cela implique généralement l'application de techniques d'apprentissage automatique (ML), dont le but est d'apprendre à catégoriser de nouveaux éléments d'information sur la base d'informations précédemment vues qui ont été explicitement ou implicitement étiquetées comme intéressantes ou non par l'utilisateur. Compte tenu de ces éléments d'information étiquetés, les méthodes ML sont capables de générer un modèle prédictif qui aidera à décider s'il est susceptible d'intéresser l'utilisateur cible.

5.1. Représentation de l'item

Les items pouvant être recommandés à l'utilisateur sont représentés par un ensemble de caractéristiques, également appelées attributs ou propriétés. Par exemple, dans une

application de recommandation de film, les caractéristiques adoptées pour décrire un film sont : (les acteurs, les réalisateurs, les genres, le sujet, …). Lorsque chaque item est décrit par le même ensemble d'attributs et qu'il existe un ensemble connu de valeurs que les attributs peuvent prendre, l'item est représenté au moyen de données structurées. Dans ce cas, de nombreux algorithmes ML peuvent être utilisés pour apprendre un profil utilisateur (M. J. Pazzani & Billsus, 2007).

Dans la plupart des systèmes de filtrage basés sur le contenu, les descriptions d'articles sont des éléments textuels extraits de pages Web, d'e-mails, d'articles de presse ou de descriptions de produits. Contrairement aux données structurées, il n'y a pas d'attributs avec des valeurs bien définies. Les caractéristiques textuelles créent un certain nombre de complications lors de l'apprentissage d'un profil d'utilisateur, en raison de l'ambiguïté du langage naturel. Le problème est que les profils traditionnels basés sur des mots-clés sont incapables de capturer la sémantique des intérêts des utilisateurs car ils sont principalement conduits par une opération de correspondance de chaînes. Si une chaîne, ou une variante morphologique, est trouvée à la fois dans le profil et dans le document, une correspondance est établie et le document est considéré comme pertinent. La correspondance des chaînes souffre des problèmes suivants (Lops et al., 2011):

- POLYSÉMIE, la présence de plusieurs sens pour un mot ;
- SYNONYMY, plusieurs mots ayant le même sens.

Le résultat est que, en raison de la synonymie, des informations pertinentes peuvent être manquées si le profil ne contient pas les mots-clés exacts dans les documents alors que, en raison de la polysémie, des documents erronés pourraient être jugés pertinents.

L'analyse sémantique et son intégration dans les modèles de personnalisation est l'une

des approches les plus innovantes et intéressantes proposées dans la littérature pour résoudre ces problèmes. L'idée clé est l'adoption de bases de connaissances, telles que des lexiques ou des ontologies, pour annoter des éléments et représenter des profils afin d'obtenir une interprétation « sémantique » des besoins d'information des utilisateurs.

5.1.1. Modèle d'espace vectoriel basé sur des mots-clés

La plupart des systèmes de recommandation basés sur le contenu utilisent des modèles de récupération relativement simples, tels que la correspondance de mots clés ou le modèle d'espace vectoriel (VSM) avec une pondération TF-IDF de base. VSM est une représentation spatiale de documents texte. Dans ce modèle, chaque document est représenté par un vecteur dans un espace à n dimensions, où chaque dimension correspond à un terme du vocabulaire global d'une collection de documents donnée. Formellement, chaque document est représenté comme un vecteur de poids de terme, où chaque poids indique le degré d'association entre le document et le terme. Soit le dictionnaire, c'est-à-dire l'ensemble des mots du corpus. T est obtenu en appliquant D = $\{d_1,d_2, ...,d_N\}$ désigne un ensemble de documents ou corpus, et T = $\{t_1,t_2, ...,t_n\}$ désigne des opérations standard de traitement du langage naturel, telles que tokenisation, suppression des mots vides et radicalisation (Baeza-Yates & Ribeiro-Neto, 1999). Chaque document dj est représenté comme un vecteur dans un terme tk dans le document dj. espace vectoriel à n dimensions, donc dj = {w1j,w2j, ...,wnj}, où wk,j est le poids du terme tk dans le document dj.

La représentation des documents dans le VSM soulève deux problèmes : la pondération des termes et la mesure de la similarité des vecteurs de caractéristiques. Le schéma de pondération des termes le plus couramment utilisé, la pondération TF-IDF (Term Frequency-Inverse Document Frequency), est basé sur des observations empiriques

concernant le texte (Salter & Antonopoulos, 2006) :

- les termes rares ne sont pas moins pertinents que les termes fréquents (hypothèse IDF);
- les occurrences multiples d'un terme dans un document ne sont pas moins pertinentes que les occurrences uniques (hypothèse TF) ;
- les documents longs ne sont pas préférés aux documents courts (hypothèse de normalisation).

En d'autres termes, les termes qui apparaissent fréquemment dans un document (TF = fréquence-terme), mais rarement dans le reste du corpus (IDF = fréquence-document inverse), sont plus susceptibles d'être pertinents pour le sujet du document. De plus, la normalisation des vecteurs de poids résultants empêche les documents plus longs d'avoir de meilleures chances d'être récupérés. Ces hypothèses sont bien illustrées par la fonction TF-IDF :

$$\boldsymbol{TF - IDF(t_k, d_j) = \frac{f_{k,j}}{max_z f_{z,j}} * log \frac{N}{n_k}} \qquad (9)$$

où $f_{k,j}$ dénote le nombre d'occurrences du terme t_k dans le document d_j et $max_z f_{z,j}$ représente la maximum des fréquences $f_{z,j}$ de tous les termes présents dans le document dj. N représente le nombre de documents dans le corpus et nk représente le nombre de documents contenant au moins une occurrence du terme t_k.

De plus, la normalisation des vecteurs est requise. Le but est que les documents soient représentés par des vecteurs de longueur égale, les poids obtenus par l'équation précédente sont généralement normalisés par la normalisation en cosinus :

$$\boldsymbol{w_{k,j} = \frac{TF-IDF(t_k,d_j)}{\sqrt{\sum_{s=1}^{|T|} TF-IDF(t_s,d_j)^2}}} \qquad (10)$$

Comme indiqué précédemment, une mesure de similitude est nécessaire pour déterminer la proximité entre deux documents. De nombreuses mesures de similarité ont

été dérivées pour décrire la proximité de deux vecteurs ; parmi ces mesures, la similarité en cosinus est la plus largement utilisée qui est calculé selon la formule suivante :

$$\boldsymbol{sim}(\boldsymbol{d_i}, \boldsymbol{d_j}) = \frac{\sum_k \boldsymbol{w_{ki}} * \boldsymbol{w_{kj}}}{\sqrt{\sum_k \boldsymbol{w_{ki}}^2} * \sqrt{\sum_k \boldsymbol{w_{kj}}^2}} \qquad (11)$$

Dans les systèmes de recommandation basés sur le contenu reposant sur VSM, les profils d'utilisateurs et les éléments sont représentés sous forme de vecteurs de termes pondérés. Les prédictions de l'intérêt d'un utilisateur pour un élément particulier peuvent être dérivées en calculant la similarité en cosinus.

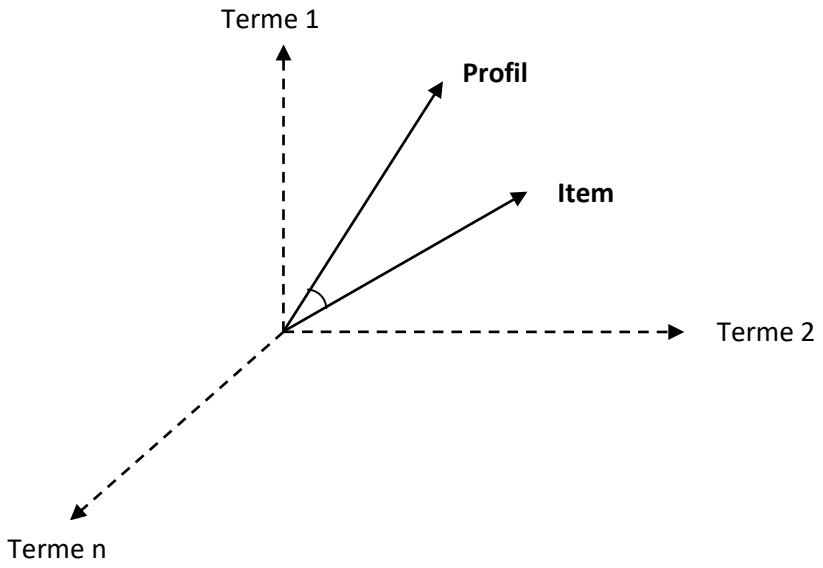


Figure 6: Illustration des systèmes de recommandation basés sur le contenu reposant sur VSM

5.1.2. Analyse sémantique basant sur l'ontologie

L'analyse sémantique permet d'apprendre des profils plus précis qui contiennent des références à des concepts définis dans des bases de connaissances externes. La principale motivation de cette approche est le défi de fournir un système de recommandation avec les connaissances culturelles et linguistiques de base qui caractérise la capacité d'interpréter

les documents en langue naturelle et le raisonnement sur leur contenu.

Dans cette section, nous présentons un examen des principales stratégies adoptées pour introduire une certaine sémantique dans le processus de recommandation. La description de ces stratégies s'effectue en tenant compte de plusieurs critères :

• le type de source de connaissances en cause (p. ex., lexique, ontologie, etc.);

• les techniques adoptées pour l'annotation ou la représentation des éléments;

• le type de contenu inclus dans le profil de l'utilisateur;

• la stratégie d'appariement des profils d'articles.

On trouve plusieurs exemples des systèmes de recommandation à base de contenu qui introduisent la sémantique pour fournir des recommandations à savoir :

- SiteIF (Magnini & Strapparava, 2001) est un agent personnel pour un site Web d'information multilingue.

- ITR (ITem Recommender) est un système capable de fournir des recommandations pour des articles dans plusieurs domaines (p. ex., films, musique, livres), à condition que les descriptions des articles soient disponibles sous forme de documents textuels (p. ex., résumés de parcelles, revues, courts résumés) (Degemmis et al., 2007).

- SEWeP (Semantic Enhancement for Web Personalization) (Eirinaki et al., 2003) est un système de personnalisation Web qui utilise à la fois les journaux d'utilisation et la sémantique du contenu d'un site Web afin de le personnaliser.

- Quickstep (Middleton et al., 2004) est un système de recommandation de documents de recherche universitaires en ligne.

- News@hand (Cantador et al., 2008) est un système qui adopte une représentation ontologique des caractéristiques des articles et des préférences des utilisateurs pour recommander des nouvelles.

- Le système JUMP (Iaquinta et al., 2008)(Basile et al., 2008) est capable de fournir intelligemment des informations contextuelles et personnalisées aux travailleurs du savoir agissant dans leur environnement de travail quotidien sur des tâches non routinières.

5.1.3. Analyse sémantique basant sur l'encyclopédie

Le bon sens et les connaissances spécifiques au domaine peuvent être utiles pour améliorer l'efficacité des techniques de traitement du langage naturel en générant des caractéristiques plus informatives que le simple sac de mots. Le processus d'apprentissage des profils d'utilisateurs pourrait bénéficier de l'infusion de connaissances exogènes (fournies de l'extérieur), en ce qui concerne l'utilisation classique des connaissances endogènes (extraites des documents eux-mêmes). De nombreuses sources de connaissances mondiales sont devenues disponibles ces dernières années.

Exemples de bases de connaissances à usage général : Open Directory Project (ODP), Yahoo! Web Directory et Wikipedia.

Dans ce qui suit, nous donnons un bref aperçu des nouvelles méthodes pour générer de nouvelles fonctionnalités avancées en utilisant la connaissance du monde, même si ces méthodes ne sont pas encore utilisées dans le contexte de l'apprentissage des profils d'utilisateurs.

L'analyse sémantique explicite (ESA) (Gabrilovich & Markovitch, 2006) (Gabrilovich & Markovitch, 2006) est une technique capable de fournir une représentation

sémantique fine de textes en langage naturel dans un espace de grande dimension de concepts naturels (et aussi compréhensibles) dérivés de Wikipedia .

Une autre approche intéressante pour ajouter de la sémantique au texte est proposée par Wikify! système (Mihalcea & Csomai, 2007)(Mihalcea & Csomai, 2007), qui a la capacité d'identifier des concepts importants dans un texte (extraction de mot-clé), puis lier ces concepts aux pages Wikipédia correspondantes (désambiguation de sens de mot).

Dans (Lees-Miller et al., 2008), Wikipedia est utilisé pour estimer la similarité entre les films, afin de fournir des prédictions plus précises pour le concours Netflix Prize. Plus précisément, le contenu et la structure des hyperliens des articles de Wikipédia sont exploités pour identifier les similitudes entre les films.

Dans (Smirnov & Krizhanovsky, 2008), une approche assez complexe, mais pas encore complète, pour le filtrage des flux RSS et des e-mails est présentée. Plus précisément, les auteurs présentent une approche utilisant Wikipedia pour générer automatiquement le profil utilisateur à partir de la collection de documents de l'utilisateur.

Dans (Semeraro et al., 2009), une approche différente pour exploiter Wikipedia dans l'étape d'analyse de contenu est présentée. Plus précisément, l'idée est de fournir un processus d'infusion de connaissances dans les systèmes de recommandation fondés sur le contenu, afin de leur fournir les connaissances culturelles de base qui, nous l'espérons, permet une analyse de contenu plus précise que les approches classiques fondées sur les mots.

5.2. Méthodes d'apprentissage des profils d'utilisateurs

Les techniques d'apprentissage automatique, généralement utilisées pour induire des profils basés sur le contenu, sont bien adaptées à la catégorisation du texte (Sebastiani,

2002). Dans une approche d'apprentissage automatique de la catégorisation de texte, un processus inductif construit automatiquement un classificateur de texte en apprenant à partir d'un ensemble de documents de formation (documents étiquetés avec les catégories auxquelles ils appartiennent) les caractéristiques des catégories.

Le problème de l'apprentissage des profils d'utilisateurs peut être considéré comme une tâche de catégorisation de texte binaire : chaque document doit être classé comme intéressant ou non par rapport aux préférences de l'utilisateur. Par conséquent, l'ensemble des catégories est C = {c+, c }, où c+ est la classe positive (utilisateurs-aiment) et c le négatif (utilisateurs-détestent).

Dans les sections suivantes, nous passons en revue les algorithmes d'apprentissage les plus utilisés dans les systèmes de recommandation de contenu. Ils sont capables d'apprendre une fonction qui modèle les intérêts de chaque utilisateur. Ces méthodes obligent généralement les utilisateurs à étiqueter les documents en attribuant un score de pertinence et à déduire automatiquement les profils exploités dans le processus de filtrage pour classer les documents en fonction des préférences de l'utilisateur.

5.2.1. Méthodes probabilistes et Naïve bayésienne

Naïve Bayes est une approche probabiliste de l'apprentissage inductif, et appartient à la classe générale des classificateurs bayésiens. Ces approches génèrent un modèle probabiliste basé sur des données précédemment observées. Le modèle estime la probabilité a posteriori, P(c|d), du document d appartenant à la classe c. Cette estimation est basée sur la probabilité a priori, P(c), la probabilité d'observer un document en classe c, P(d|c), la probabilité d'observer le document d donné c, et P(d), la probabilité d'observer l'instance d. En utilisant ces probabilités, le théorème de Bayésienne est appliqué pour calculer P(c|d):

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)} \quad (12)$$

Pour classer le document d, on choisit la classe la plus probable :

$$C = argmaxcj \frac{P(cj)P(d|cj)}{P(d)} \quad (13)$$

P(d) est généralement supprimé car il est égal pour tous cj. Comme nous ne connaissons pas la valeur pour P(d|c) et P(c), nous les estimons en observant les données de formation. Cependant, estimer P(d|c) de cette façon est problématique, car il est très peu probable de voir le même document plus d'une fois : les données observées ne sont généralement pas suffisantes pour générer de bonnes probabilités. Le classificateur naïve Bayésienne surmonte ce problème en simplifiant le modèle à travers l'hypothèse d'indépendance : tous les mots ou jetons dans le document observé d sont conditionnellement indépendants les uns des autres compte tenu de la classe. Les probabilités individuelles pour les mots dans un document sont estimées une par une plutôt que le document complet dans son ensemble. L'hypothèse d'indépendance conditionnelle est clairement violée dans les données du monde réel, cependant, malgré ces violations, empiriquement le classificateur naïve Bayésienne fait un bon travail dans la classification des documents texte (Lewis, 1994)(Billsus & Pazzani, 1997).

Il existe deux modèles de travail couramment utilisés du classificateur naïve Bayésienne, le modèle d'événement multi-varié Bernoulli et le modèle d'événement multinomial (Maertens et al., 2017). Les deux modèles traitent un document comme un vecteur de valeurs sur le vocabulaire du corpus, où chaque entrée dans le vecteur représente

si un mot s'est produit dans le document, donc les deux modèles perdent des informations sur l'ordre des mots. Le modèle d'événement multivarié Bernoulli code chaque mot comme un attribut binaire, c'est à dire, si un mot est apparu ou non, tandis que le modèle d'événement multinomial compte combien de fois le mot est apparu dans le document. Empiriquement, il a été démontré que la formulation multinomiale naïve Bayésienne surperforme le modèle multivarié de Bernoulli. Cet effet est particulièrement marqué pour les grands vocabulaires (Maertens et al., 2017).

Le classificateur naïve Bayésienne a été utilisé dans plusieurs systèmes de recommandation basés sur le contenu, tels que Syskill & Webert (Science & Wnek, 1997)(M. Pazzani et al., 1996), NewsDude (Billsus & Pazzani, 1999), Daily Learner (Billsus Daniel & Pazzani Michael J., 2002), LIBRA (Mooney & Roy, 2000) et ITR (Degemmis et al., 2007).

5.2.2. Pertinence Feedback et algorithme de Rocchio

La rétroaction sur la pertinence est une technique adoptée dans la recherche documentaire qui aide les utilisateurs à affiner progressivement les requêtes en fonction des résultats de recherche précédents. Il consiste à ce que les utilisateurs réintègrent les décisions du système sur la pertinence des documents récupérés par rapport à leurs besoins d'information.

La rétroaction sur la pertinence et son adaptation à la catégorisation des textes, la formule bien connue de Rocchio (Lv & Zhai, 2009), sont couramment adoptées par les systèmes de recommandation fondés sur le contenu. Le principe général est de permettre aux utilisateurs d'évaluer les documents suggérés par le système de recommandation en

fonction de leurs besoins d'information. Cette forme de rétroaction peut ensuite être utilisée pour affiner progressivement le profil de l'utilisateur ou pour former l'algorithme d'apprentissage qui infère le profil de l'utilisateur en tant que classificateur.

Certains classificateurs linéaires consistent en un profil explicite (ou document prototypique) de la catégorie (Sebastiani, 2002). La méthode de Rocchio est utilisée pour induire des classificateurs linéaires de type profil. Cet algorithme représente les documents comme des vecteurs, de sorte que les documents ayant un contenu similaire ont des vecteurs similaires. Chaque composante d'un tel vecteur correspond à un terme du document, généralement un mot. La pondération de chaque composante est calculée à l'aide du schéma de pondération des échéances de la FO-FIL. L'apprentissage est réalisé en combinant les vecteurs de documents (des exemples positifs et négatifs) en un vecteur prototype pour chaque classe dans l'ensemble des classes C. Pour classer un nouveau document d, la similarité entre les vecteurs prototypes et le vecteur de document correspondant représentant d est calculée pour chaque classe (par exemple en utilisant la mesure de similarité cosinus), puis d est attribué à la classe dont le vecteur de document a la similarité la plus élevée.

5.2.3. Autre Méthodes

D'autres algorithmes d'apprentissage ont été utilisés dans les systèmes de recommandation basés sur le contenu. Une brève description des algorithmes les plus importants suit. Un examen approfondi est présenté dans (Sebastiani, 2002)(Montaner et al., 2003)(M. J. Pazzani & Billsus, 2007).

Les arbres de décision sont des arbres dans lesquels les nœuds internes sont étiquetés

par des termes, les branches qui en sortent sont étiquetées par des tests sur le poids que le terme a dans le document de test, et les feuilles sont étiquetées par catégories. Les arbres de décision sont appris en répartissant récursivement les données de formation, c'est-à-dire les documents texte, en sous-groupes, jusqu'à ce que ces sous-groupes ne contiennent que des instances d'une seule classe. Le test pour les données de partitionnement est exécuté sur les poids que les termes étiquetant les nœuds internes ont dans le document. Le choix du terme sur lequel opérer la partition se fait généralement selon un critère de gain d'information ou d'entropie (Yiming & Jan O., 1997). Les arbres de décision sont utilisés dans le système de recommandation Syskill & Webert (Science & Wnek, 1997)(M. Pazzani et al., 1996).

Les classificateurs de règles de décision sont similaires aux arbres de décision, car ils fonctionnent de manière similaire à l'approche récursive de partage des données décrite ci-dessus. Un avantage des apprenants de règles est qu'ils ont tendance à générer des classificateurs plus compacts que les apprenants des arbres de décision. Les méthodes d'apprentissage des règles tentent habituellement de choisir parmi toutes les règles possibles (c.-à-d. les règles qui classent correctement tous les exemples de formation) la meilleure selon un critère de minimalité.

Les algorithmes voisins les plus proches stockent simplement les données d'entraînement en mémoire, et classent un nouvel élément invisible en le comparant à tous les éléments stockés en utilisant une fonction de similarité. Les items « voisin le plus proche » ou « voisin le plus proche » sont déterminés, et l'étiquette de classe pour l'item non classifié est dérivée des étiquettes de classe des voisins les plus proches. Une fonction de similarité est nécessaire, par exemple la mesure de similarité cosinus est adoptée lorsque

les items sont représentés à l'aide de la VSM. Les algorithmes voisins les plus proches sont très efficaces, bien que l'inconvénient le plus important soit leur inefficacité au moment de la classification, car ils n'ont pas une vraie phase de formation et donc reporter tout le calcul au temps de classification. Daily Learner (Billsus Daniel & Pazzani Michael J., 2002) et Quickstep (Kobayashi, 2005) utilisent l'algorithme voisin le plus proche pour créer un modèle de l'intérêt à court terme de l'utilisateur et pour associer des annotations sémantiques d'articles avec des noms de classe dans l'ontologie, respectivement.

6. Approche hybride

Les systèmes de recommandation hybrides combinent plusieurs systèmes de recommandation. Cependant, dans un système hybride, plusieurs types différents de systèmes de recommandation sont combinés. Un aperçu complet des différentes techniques d'hybridation est donné par Burke (Burke, 2002). Comme l'a conclu Burke, étant donné une certaine hybridation.

Afin d'améliorer les performances tout en minimisant les inconvénients de chaque technique de recommandation, les systèmes de recommandation hybrides intègrent deux ou plus des techniques. Pour contourner le problème de montée en puissance, le filtrage collaboratif est le plus souvent utilisé en conjonction avec une autre méthode. Une bonne revue des approches hybrides peut être trouvée dans (Adomavicius & Tuzhilin, 2005)(Burke, 2002). Burke classe les techniques d'hybridation en sept classes.

6.1. Système de recommandation hybride pondéré

Le système de recommandation hybride pondéré (weighted) : chacune des approches

de recommandation fait des prédictions qui sont ensuite combinées en une seule prédiction.

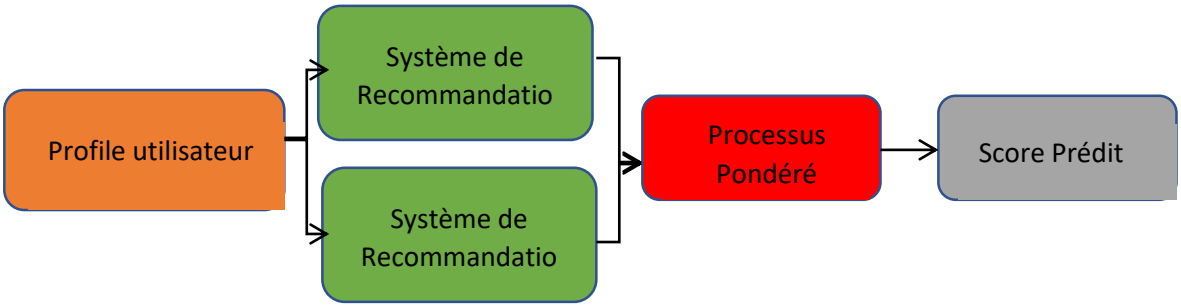


Figure 7: Exemple d'un système de recommandation hybride pondéré

Dans le système de recommandation pondéré, nous pouvons définir quelques modèles capables de bien interpréter l'ensemble de données. Le système de recommandation pondéré prendra les sorties de chacun des modèles et combinera le résultat dans des pondérations statiques, dont le poids ne change pas à travers le train et l'ensemble de test.

Par exemple, nous pouvons combiner un modèle basé sur le contenu et un modèle de filtrage collaboratif élément-élément, et chacun prend un poids de 50 % dans la prédiction finale.

L'avantage de l'utilisation de l'hybride pondéré est que nous intégrons plusieurs modèles pour prendre en charge l'ensemble de données sur le processus de recommandation de manière linéaire.

6.2. Système de recommandation hybride commutation

Le système de recommandation hybride commutation (switching) : l'une des techniques de recommandation est sélectionnée pour faire la prédiction lorsque certains critères sont remplis.

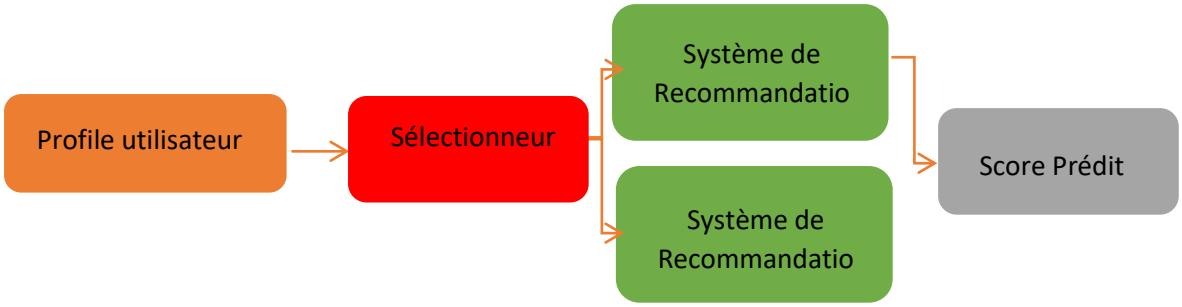


Figure 8: Exemple d'un système de recommandation hybride commutatif

L'hybride de commutation sélectionne un système de recommandation unique en fonction de la situation. Le modèle est utilisé pour être construit pour l'ensemble de données sensibles au niveau de l'élément, nous devons définir les critères du sélecteur de recommandation en fonction du profil de l'utilisateur ou d'autres fonctionnalités.

L'approche hybride de commutation introduit une couche supplémentaire sur le modèle de recommandation, qui sélectionne le modèle approprié à utiliser. Le système de recommandation est sensible aux forces et aux faiblesses du modèle de recommandation constituant.

6.3. Système de recommandation hybride Mixte

Le système de recommandation hybride Mixte (mixed) : les prédictions de chacune des techniques de recommandation sont présentées à l'utilisateur.

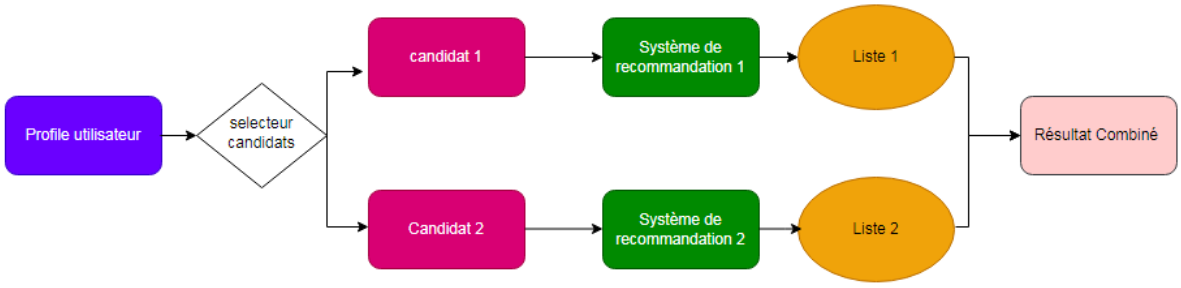


Figure 9: Illustration des systèmes de recommandation mixte

L'approche hybride mixte prend d'abord le profil et les fonctionnalités de l'utilisateur pour générer différents ensembles de données candidats. Le système de recommandation entre différents ensembles de candidats dans le modèle de recommandation en conséquence, et combine la prédiction pour produire la recommandation de résultat.

Le système de recommandation hybride mixte est capable de faire un grand nombre de recommandations simultanément et d'adapter l'ensemble de données partiel au modèle approprié afin d'avoir de meilleures performances.

6.4. Système de recommandation hybride Combinaison de caractéristiques

Le système de recommandation hybride Combinaison de caractéristiques (feature combination) : un seul algorithme de prédiction est fourni avec des caractéristiques de différentes techniques de recommandation.

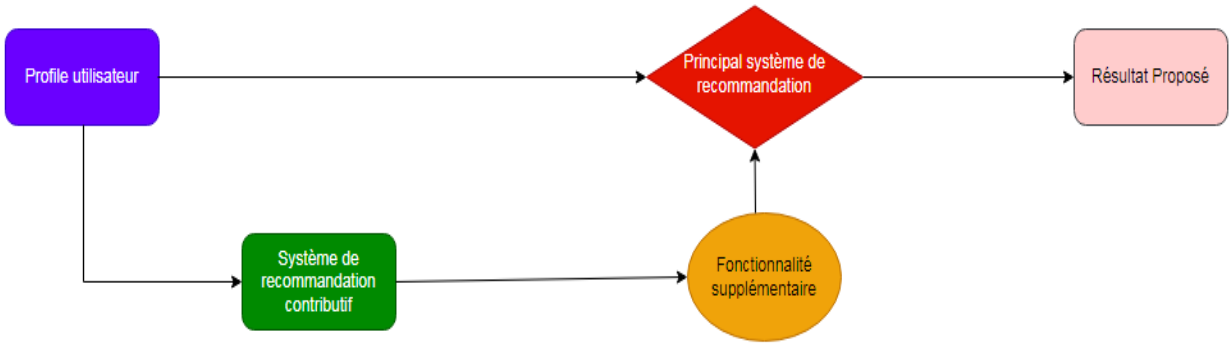


Figure 10: Exemple des systèmes de recommandation hybride combinaison de caractéristiques

Dans l'hybride de combinaison de fonctionnalités, nous ajoutons un modèle de recommandation contributif virtuel au système, qui fonctionne comme une ingénierie de fonctionnalités vers l'ensemble de données de profil utilisateur d'origine.

Par exemple, nous pouvons injecter des fonctionnalités d'un modèle de recommandation collaboratif dans un modèle de recommandation basé sur le contenu. Le modèle hybride est capable de considérer les données collaboratives du sous-système en s'appuyant exclusivement sur un modèle.

6.5. Système de recommandation hybride cascade

Le système de recommandation hybride cascade (cascade) : le résultat d'une technique de recommandation est affiné par une autre.

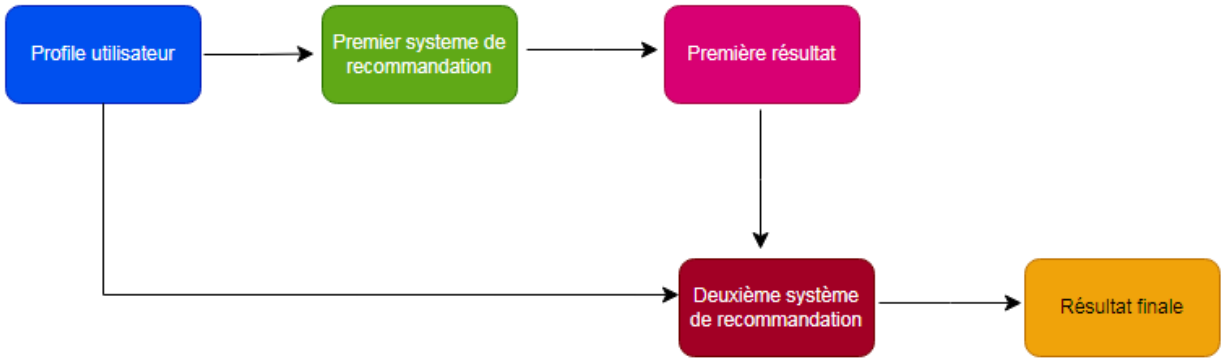


Figure 11: Exemple des systèmes de recommandation hybride en cascade

Système de recommandation hybride en cascade définit un système de recommandation de structure hiérarchique stricte, de sorte que le système de recommandation principal produise le résultat principal, et nous utilisons le modèle secondaire pour résoudre certains problèmes mineurs du résultat principal, comme la rupture d'égalité dans la notation.

En pratique, la plupart des ensembles de données sont clairsemés, le modèle de recommandation secondaire peut être efficace contre le problème de score égal ou le problème de données manquantes.

6.6. Système de recommandation hybride augmentation de caractéristiques

Le système de recommandation hybride augmentation de caractéristiques (feature augmentation) : la sortie d'une technique de recommandation est transmise à une autre.



Figure 12: Exemple des systèmes de recommandation hybride augmentation de caractéristiques

Un modèle de recommandation contributif est utilisé pour générer une évaluation ou une classification du profil utilisateur/élément, qui est en outre utilisée dans le système de recommandation principal pour produire le résultat prédit final.

L'hybride d'augmentation de fonctionnalités est capable d'améliorer les performances du système central sans modifier le modèle de recommandation principal. Par exemple, en utilisant la règle d'association, nous sommes en mesure d'améliorer l'ensemble de données de profil utilisateur. Avec l'ensemble de données augmenté, les performances du modèle de recommandation basé sur le contenu seront améliorées.

6.7. Système de recommandation hybride méta-niveau

Le système de recommandation hybride méta-niveau (Meta-level) : l'ensemble du modèle produit par une technique de recommandation est utilisé par une autre.

L'hybride de méta-niveau est similaire à l'hybride d'augmentation de fonctionnalités, de sorte que le modèle de contribution fournit un ensemble de données augmenté au modèle de recommandation principal. Différent de l'hybride d'augmentation de fonctionnalités, le méta-niveau remplace l'ensemble de données d'origine par un modèle appris du modèle de contribution comme entrée du modèle de recommandation principal.

7. Méthodologies d'évaluation

Les méthodologies d'évaluation sont des approches systématiques et structurées pour évaluer et mesurer la performance, l'efficacité, la qualité ou d'autres aspects d'un système, d'un processus, d'un produit ou d'un service. Elles sont largement utilisées dans de nombreux domaines pour garantir que les objectifs sont atteints, que les normes sont respectées et que des améliorations sont apportées lorsque cela est nécessaire.

L'évaluation des systèmes de recommandation est une tâche difficile car les systèmes de recommandation doivent fournir des recommandations utiles qui sont pertinentes pour l'utilisateur final. Les méthodes d'évaluation les plus courantes sont les évaluations hors ligne, les évaluations en ligne et les études des utilisateurs (user studies)(Ricci et al., 2015).

7.1. Evaluations hors ligne (Offline Evaluation)

La méthode d'évaluation hors ligne dans le domaine des systèmes de recommandation est un moyen plus pratique d'évaluer le système de recommandation. Dans cette méthode, l'ensemble de données qui contient les informations des utilisateurs, les items et les évaluations sont divisés en ensembles d'apprentissage et de validation pour former un modèle sur l'ensemble d'apprentissage et tester le modèle sur l'ensemble de validation. Tandis que les performances du système sont ensuite évaluées sur l'ensemble de validation avec différentes techniques d'évaluation. La méthode d'évaluation hors ligne est la plus pratique à utiliser car elle permet de comparer les différents algorithmes de recommandation.

Étant donné que la méthode fonctionne sur les données et le comportement des

utilisateurs collectés dans le passé, la méthode d'évaluation hors ligne ne traite pas les utilisateurs actuels. Cependant, la méthode est utile lorsque les expériences doivent être répétées de nombreuses fois, en particulier dans un environnement de recherche.

Les paramètres de mesure de l'exactitude des systèmes de filtrage des recommandations sont divisés en critères d'exactitude statistique et d'aide à la décision (Sarwar et al., 2001b).

7.1.1. Mesures d'exactitude statistique

Les mesures d'exactitude statistique évaluent l'exactitude d'une technique de filtrage en comparant les cotes prévues directement avec la cote réelle de l'utilisateur. L'erreur absolue moyenne (MAE) (Breese et al., 1998)(Jonathan L. Herlocker, Joseph A. Konstan, Al Borchers, 1999), l'erreur quadratique moyenne (RMSE) et la corrélation sont habituellement utilisées comme mesures d'exactitude statistique. Le MAE est le plus populaire et le plus utilisé; il s'agit d'une mesure de l'écart entre la recommandation et la valeur spécifique de l'utilisateur. Il est calculé comme suit (Claypool et al., 1999) :

7.1.1.1. Erreur Absolue Moyenne (MAE)

L'erreur Absolue Moyenne (MAE), en anglais Mean Absolute Error, est une mesure couramment utilisée pour évaluer la précision d'un modèle de prédiction ou de régression. Elle mesure la moyenne des valeurs absolues des erreurs entre les prédictions d'un modèle et les valeurs réelles de l'ensemble de données. En d'autres termes, la MAE quantifie à quel point les prédictions d'un modèle sont éloignées des valeurs réelles, sans tenir compte du signe de l'erreur.

La formule mathématique de la MAE est la suivante :

$$MAE = \frac{\sum_{(s,i)}|p_{s,i} - r_{s,i}|}{N} \quad (14)$$

où $p_{s,i}$est la note prévue pour l'utilisateur $\boldsymbol{u}$ sur l'item $\boldsymbol{i}$, $r_{s,i}$est la note réelle et N est le nombre total de notes sur le jeu d'éléments.

La MAE mesure l'erreur moyenne entre les valeurs prédites par le modèle et les valeurs réelles, en prenant la valeur absolue de chaque différence. Plus la MAE est faible, plus le modèle est précis. Une MAE égale à zéro signifierait que le modèle fait des prédictions parfaites. Cependant, cela est rare dans la pratique, et la MAE est souvent utilisée pour évaluer la qualité des modèles de prédiction.

7.1.1.2. Racine d'erreur quadratique moyenne (RMSE)

La racine d'erreur quadratique moyenne (Root Mean Squared Error en anglais) est utilisée pour calculer la précision des recommandations en fonction de la prédiction des notes dans l'ensemble de validation à partir de l'ensemble de données choisi (Panniello et al., 2014). Il mesure l'ampleur moyenne des erreurs dans les valeurs prédites. C'est la distance moyenne d'un point de données à partir de la droite d'ajustement. Étant une mesure quadratique, l'erreur quadratique moyenne (RMSE) est la plus affectée par les erreurs importantes, elle est donc utile lorsque les erreurs importantes sont particulièrement indésirables. Le système de recommandation produit les évaluations prévues pour l'utilisateur et les éléments et les évaluations données. La racine carrée moyenne entre les notes prédites et données peut être calculée par l'équation suivante :

$$RMSE = \sqrt{\frac{\sum_{(s,i)}(p_{s,i} - r_{s,i})2}{N}} \quad (15)$$

où $p_{s,i}$est la note prévue pour l'utilisateur u sur l'item i, $r_{s,i}$est la note réelle et N est le nombre total de notes sur le jeu d'éléments

7.1.1.3. Erreur quadratique moyenne (MSE)

L'erreur quadratique moyenne, ou Mean Squared Error (MSE) en anglais, est une mesure couramment utilisée pour évaluer la qualité d'un modèle de prédiction, en particulier dans le domaine de la régression statistique. Il quantifie la moyenne des carrés des écarts entre les valeurs prédites par le modèle et les valeurs réelles observées.

$$MSE = \frac{1}{n}\sum(s,i)\left(p_{s,i} - r_{s,i}\right)2 \quad (16)$$

Plus le MSE est bas, plus le modèle est précis. Un MSE de zéro signifierait que le modèle prédit parfaitement les valeurs réelles, mais dans la pratique, cela est rarement atteint. Le MSE est une métrique utile pour comparer différents modèles ou ajuster les paramètres d'un modèle, car il permet de quantifier la précision des prédictions.

7.1.1.4. Erreur Absolue Moyenne Normalisée (NMAE)

L'erreur Absolue Moyenne Normalisée La (NMAE), ou Normalized Mean Absolute Error en anglais, est une mesure d'évaluation couramment utilisée pour évaluer la performance des modèles de prédiction, en particulier dans le domaine de la régression. La NMAE est une version normalisée de l'erreur absolue moyenne (MAE) qui tient compte de la plage des valeurs de la variable cible.

La formule mathématique de la NMAE est la suivante :

$$NMAE = \frac{\sum_{(s,i)}(|p_{s,i} - r_{s,i}|)/(p_{Max} - p_{Min})}{N} \quad (17)$$

La NMAE normalise l'erreur absolue moyenne en divisant par la plage des valeurs de la variable cible. Cela permet de comparer les performances des modèles sur des ensembles de données avec des plages de valeurs différentes. Une NMAE plus faible indique une meilleure précision du modèle, mais il n'y a pas de seuil standard pour évaluer la qualité d'un modèle à partir de la NMAE, cela dépend du contexte spécifique de l'application.

7.1.2. Mesures d'aide à la décision

Les performances des recommandations, en particulier dans des domaines unaires tels que les historiques d'achat, peuvent également être évaluées dans le cadre de Précision-Rappel (Precision-Recall en anglais) de la recherche d'informations dérivée de la théorie statistique de la décision (Su & Khoshgoftaar, 2009). Ce cadre examine la capacité d'un système de recherche à identifier avec précision les ressources pertinentes pour une requête, en mesurant séparément sa capacité à trouver tous les éléments pertinents et à éviter de trouver des items non pertinents. La figure 13 montre la matrice de confusion, illustrant la relation entre ces concepts. Les recommandations sont évalués avec des métriques de récupération en opérant dans un domaine d'achat unaire, en retenant certains articles hors du profil de l'utilisateur et en considérant les articles achetés retenus comme pertinents pour une requête consistant en leur profil restant (Breese et al., 1998)(Li et al., 2017)(Sarwar et al., 2000). Ces métriques peuvent également être appliquées dans des domaines d'évaluation non unaires en utilisant un seuil d'évaluation pour faire la distinction entre les éléments

appréciés et détestés.

Pour être appliquées de manière robuste, toutes ces mesures nécessitent un corpus entièrement codé où chaque élément est connu pour être soit pertinent, soit non pertinent pour la requête considérée. Dans l'évaluation des recommandations, cependant, ces données ne sont généralement pas disponibles - si un article n'a pas été acheté, cela peut être parce que l'utilisateur n'aime pas l'article ou parce qu'il n'en était pas conscient. Si le recommandeur renvoie un élément en dehors de l'ensemble cible retenu, il peut toujours s'agir d'un élément pertinent. Des précautions doivent donc être prises lors de la conception et de l'évaluation des évaluations de Précision- Rappel des systèmes de recommandation ; néanmoins, il peut s'agir d'un cadre utile pour comprendre les performances de l'outil de recommandation.

	Relevant	Irrelevant
Retrieved	TP	FP
Not retrieved	FN	TN

Figure 13: Matrice de confusion d'extraction.

$$\text{Précision} = \frac{TP}{TP + FP} \quad (18)$$

$$\text{Rappel} = \frac{TP}{TP + FN} \quad (19)$$

Les métriques de base dans les évaluations de précision-rappel sont la précision P, la fraction d'articles retournés par le recommandeur qui sont achetés (18) et le rappel R, la fraction d'articles achetés retournés par le recommandeur (19) (Salton, 1992). La figure 9 montre comment ces métriques sont liées à la matrice de confusion. Étant donné que les systèmes de recommandation et de récupération renvoient généralement des listes classées,

celles-ci sont généralement calculées pour une longueur de liste de recommandation fixe N ou parfois pour un score seuil. Pour un système donné, la précision et le rappel sont inversement liés et dépendent de N , de sorte que la comparaison des résultats entre les recommandations peut être fastidieuse (Li et al., 2017) ; cependant, la combinaison peut fournir plus d'informations sur le comportement particulier d'un algorithme que les mesures à un seul nombre. Les différentes tâches de l'utilisateur nécessitent des compromis différents entre précision et rappel - un utilisateur à la recherche d'une recommandation de film ne se soucie probablement que d'obtenir un bon film (haute précision), tandis qu'un avocat à la recherche d'un précédent juridique doit trouver tous les cas pertinents (haut rappel). (Li et al., 2017) classent ces deux besoins disparates en « Find Good Items » et « Find All Good Items ».

La Précision et le Rappel peuvent être simplifiés en une seule métrique, la métrique F1 (Van Rijsbergen, 1979) (Yang & Liu, 1999), qui a été utilisée pour évaluer les systèmes de recommandation (Sarwar et al., 2000) :

$$F1 = \frac{2PR}{P + R} \quad (20)$$

La F1 allie précision et Rappel avec un poids égal. Le nombre résultant facilite la comparaison entre les algorithmes et entre les ensembles de données, mais ne facilite pas les comparaisons plus nuancées où l'une des deux mesures est plus importante que l'autre.

La précision et le rappel ignorent également le quadrant inférieur droit de la matrice de confusion - les vrais négatifs, ces éléments correctement identifiés comme non pertinents ou détestés. Afin d'incorporer ce quadrant dans une métrique, une paire similaire de

métriques appelée sensibilité et spécificité peut être utilisée (Landgrebe et al., 2006). Les équations (20) et (21) montrent ces métriques. La sensibilité équivaut au rappel (le taux de vrais positifs) tandis que la spécificité mesure la fraction d'items non pertinents correctement rejetés (le taux de vrais négatifs).

$$\boldsymbol{Sensibilité} = \frac{TP}{TP + FN} \quad (21)$$

$$\boldsymbol{Spécificité} = \frac{TN}{TP + FN} \quad (22)$$

Une méthode de comparaison associée mais plus complète est la caractéristique de fonctionnement relatif (ROC, également appelée caractéristique de fonctionnement du récepteur) (Van Rijsbergen, 1979).

7.2. Evaluation en ligne

Les évaluations en ligne proviennent de la publicité en ligne et mesurent les taux d'acceptation des recommandations dans les systèmes de recommandation du monde réel. Les taux d'acceptation sont généralement mesurés par le taux de clics (CTR), c.-à-d. le rapport entre les recommandations cliquées et les recommandations affichées. Par exemple, si un système de recommandation affiche 10000 recommandations et 120 sont cliquées, CTR est de 1,2 %. D'autres mesures comprennent le ratio des items téléchargés ou achetés. Le taux d'acceptation est généralement interprété comme une mesure implicite de la satisfaction des utilisateurs. L'hypothèse est que lorsqu'un utilisateur clique, télécharge ou achète un item recommandé, l'utilisateur a aimé la recommandation. Bien sûr, cette hypothèse n'est pas toujours fiable parce que les utilisateurs, par exemple, pourraient acheter un livre, mais après l'avoir lu encore le taux négatif. Si l'objectif du système de

recommandation est le revenu, des mesures comme le CTR peuvent être des mesures explicites de l'efficacité, à savoir lorsque l'opérateur reçoit de l'argent, p. ex., pour les clics sur les recommandations (Beel & Langer, 2015).

7.3. Etudes des utilisateurs (user studies)

Les études des utilisateurs mesurent généralement la satisfaction des utilisateurs au moyen de notes explicites. Les utilisateurs reçoivent des recommandations générées par différentes approches de recommandation, évaluent les recommandations et la communauté considère l'approche ayant la note moyenne la plus élevée comme étant la plus efficace (Adomavicius & Tuzhilin, 2005). Les participants à l'étude sont généralement invités à quantifier leur satisfaction globale à l'égard des recommandations. Une étude d'utilisateur peut également recueillir des commentaires qualitatifs, mais cela est rarement fait dans le domaine des systèmes de recommandation (document de recherche). Nous faisons une plus grande distinction entre les études de « laboratoire » et les études d'utilisateurs « réelles ». Dans les études de laboratoire, les participants savent qu'ils font partie d'une étude d'utilisateur, ce qui, ainsi que plusieurs autres facteurs, pourrait affecter leur comportement et donc les résultats de l'évaluation (Tuch et al., 2012). Dans les études du monde réel, les participants ne sont pas au courant de l'étude et évaluent les recommandations pour leur propre bénéfice, par exemple parce que le système de recommandation améliore les recommandations fondées sur les notes (c.à.d. la rétroaction sur la pertinence), ou que les notes sont nécessaires pour générer des recommandations. (c.à.d. filtrage collaboratif)

8. Conclusion

Au fil de ce chapitre, nous avons mis en évidence le filtrage collaboratif comme

une méthode solide permettant de fournir des recommandations personnalisées en se basant sur les préférences et les interactions des utilisateurs. Cependant, malgré ses avantages évidents, ce système n'est pas exempt de limites qu'il est crucial de prendre en compte.

Dans le chapitre à venir, nous approfondirons une approche spécifique visant à surmonter deux défis majeurs auxquels le filtrage collaboratif est confronté : la problématique des "moutons gris" et celle de la "matrice creuse". Nous présenterons en détail cette méthode et son application pour atténuer ces problèmes significatifs, afin d'améliorer la performance et la précision du système de recommandation.

CHAPITRE 3: FILTRAGE COLLABORATIF BASE SUR L'INFERENCE DES UTILISATEURS IMAGINAIRES

Dans ce chapitre, nous décrivons notre méthodologie de recherche qui est proposée pour le filtrage collaboratif. Nous proposons une nouvelle approche CF pour traiter le problème du mouton gris (comportement inhabituel) afin d'améliorer la précision des prédictions. Notre solution vise à déduire de nouveaux utilisateurs à partir de vrais existants dans les datasets. Cette transformation permet de créer des utilisateurs avec des préférences opposées aux vraies ce qui augmente le nombre des voisins. L'hypothèse de base est que si l'utilisateur X n'est pas similaire à l'utilisateur Y, alors l'utilisateur imaginaire ¬X est similaire à l'utilisateur Y.

1. Introduction

Au cours des dernières décennies, de nombreuses approches de système de recommandation ont été proposées et développées afin de répondre aux besoins croissants des utilisateurs et de surmonter les problèmes rencontrés dans le processus de recommandation. Selon (Adomavicius & Tuzhilin, 2005), trois principaux types de systèmes de recommandation ont été proposés dans la littérature : le Filtrage Collaboratif (CF) (El Madani El Alami et al., 2016), les systèmes de recommandation basés sur le contenu (M. J. Pazzani & Billsus, 2007) et les systèmes de recommandation hybrides (Burke, 2002)(El Madani El Alami et al., 2015).

CF reconnaît les points communs entre les utilisateurs ou entre les éléments sur la base

d'indications de pertinence (Burke, 2010). Un système de recommandation basé sur le contenu suggère les éléments qui ont des caractéristiques similaires aux éléments que l'utilisateur a aimés auparavant (Takács & Tikk, 2012). Une recommandation typique basée sur le contenu crée d'abord un profil d'utilisateur à l'aide des commentaires et des évaluations des utilisateurs sur les éléments. Un système de recommandation hybride combine plusieurs approches pour obtenir une certaine synergie entre elles (Behera et al., 2021)(Ngoupeyou Tondji, 2018).

L'approche CF est l'approche la plus utilisée dans les systèmes de recommandation en raison de leur efficacité et de leur simplicité (Ekstrand, 2011). Dans cette section, nous passons brièvement en revue les principaux algorithmes CF rapportés dans la littérature. Ces algorithmes sont basés sur une intuition simple. Ils supposent que de bonnes recommandations peuvent être dérivées d'utilisateurs partageant les mêmes intérêts et préférences. Ces préférences peuvent être exprimées de plusieurs manières, soit en utilisant des notes basées sur les intérêts des utilisateurs (Jawaheer et al., 2006), soit en déduisant du comportement des utilisateurs, en suivant leur historique d'achat et le temps passé sur le contenu Web, ce que l'on appelle le retour implicite (Y. Hu et al., 2008) . Ils peuvent être exprimés sous la forme d'une matrice appelée matrice de notation(Hernando et al., 2016). C'est la base pour créer des modèles de prédiction efficaces et des profils d'utilisateurs (Isinkaye et al., 2015a). Il existe deux approches principales de CF : basée sur la mémoire et basée sur un modèle.

- ✓ Les approches de CF basées sur la mémoire exploitent directement les préférences des utilisateurs (Gras et al., 2016) (établissant des relations similaires entre les utilisateurs ou les éléments sur la base de la matrice de notation utilisateur-élément).

Les techniques révélées dans l'approche basée sur la mémoire sont considérées comme le premier algorithme de CF (Breese et al., 1998). Les recommandations s'appliquent facilement à la matrice de notation. Les techniques utilisées dans le CF basé sur la mémoire permettent d'obtenir la similarité pour calculer la distance entre deux utilisateurs (approche basée sur l'utilisateur), ou entre deux items (approche basée sur l'item), selon les évaluations, dans la matrice de notation (Sarwar B, Karypis G, Konstan J, 2001).

- ✓ Les approches basées sur des modèles (également appelées filtrage collaboratif basé sur le voisinage) construisent ou apprennent des modèles à partir de notes collectées sur la base de techniques d'apprentissage automatique telles que les techniques de clustering (Tsai & Hung, 2012), les approches de réduction de dimensionnalité (Paterek, 2007), les machines à vecteurs de support et les réseaux de neurones (Agrawal & Agrawal, 2015).

CF s'appuie sur la communauté des utilisateurs du système. Sa principale caractéristique est l'utilisation de notes obtenues à partir d'items recommandés par les utilisateurs. Le principe est de filtrer le flux d'items évalués par la communauté des autres utilisateurs. Si un item a été jugé intéressant par un utilisateur, il sera automatiquement recommandé aux utilisateurs ayant des vues similaires dans le passé. Par conséquent, l'objectif de CF est de prédire, pour un élément non noté, l'évaluation que l'utilisateur cible pourrait attribuer, en fonction sur la corrélation entre leurs propres évaluations et les évaluations d'autres utilisateurs qui ont des intérêts et des préférences similaires.

Dans la section suivante, nous passons en revue les techniques basées sur la mémoire

en utilisant l'approche basée sur l'utilisateur (UBCF). Les K premiers utilisateurs qui ont des préférences similaires à un utilisateur donné sont appelés K voisins les plus proches (KNN). Nous utilisons des mesures de similarité pour établir des similitudes ou des corrélations entre les utilisateurs afin d'identifier le KNN. Dans KNN, la valeur de K est le nombre de voisins similaires dont nous avons besoin pour prédire les notes (Gras et al., 2016).

Malgré ses avantages, CF présente un certain nombre d'inconvénients affectant la précision des recommandations. Parmi ces inconvénients, il y a les utilisateurs de moutons gris (gray sheep), désignant des utilisateurs qui ont des préférences inhabituelles et qui ne partagent pas les mêmes préférences avec les autres utilisateurs (Burke, 2010). Ce problème rend la tâche de trouver des voisins difficile.

Ce travail vise à atténuer le problème du mouton gris et à améliorer la précision des recommandations basées sur les préférences opposées des utilisateurs. L'idée de base qui régit cela réside dans la génération d'utilisateurs imaginaires basés sur des utilisateurs dissemblables afin d'améliorer le voisinage des utilisateurs. L'hypothèse sous-jacente de notre approche est que si l'utilisateur X a une préférence opposée à l'utilisateur Y, alors l'utilisateur imaginaire ¬X a une préférence similaire à l'utilisateur Y. Notre approche augmente le nombre de voisins comparables, amplifie la densité du voisinage, et permet ensuite de construire de bonnes recommandations.

2. Problème de moutons gris (gray sheep)

Dans cette section, nous discutons du problème des utilisateurs de moutons gris et

expliquons comment ce problème est négligé dans la recherche sur les systèmes de recommandation.

Les travaux de (Claypool et al., 1999b) confirment que l'efficacité de l'algorithme CF traditionnel varie d'un utilisateur à l'autre. Il existe deux catégories principales d'utilisateurs : les moutons blancs (WS) et les moutons gris (GS). Les utilisateurs de WS ont une similarité élevée avec de nombreux autres utilisateurs (la valeur de corrélation est élevée), tandis que les utilisateurs de GS sont dissemblables ou partiellement similaires aux autres utilisateurs et ont un coefficient de corrélation plus faible avec presque tous les utilisateurs (Adamopoulos, 2014). Par conséquent, les recommandations des utilisateurs deviennent moins précises en raison des utilisateurs de GS (Srivastava et al., 2020), donc, ils ne bénéficient pas de systèmes de recommandation. Certains travaux traitent du problème des utilisateurs de moutons gris (Ghorbani & Novin, 2016)(Jindal et al., 2018)(Ghazanfar & Prugel-bennett, n.d.). (Claypool et al., 1999b) ont mis en évidence le problème de GS et mis en ligne un système de recommandation hybride pour des recommandations mises à jour. Ils ont combiné des approches de filtrage FC et basées sur le contenu, en utilisant une approche pondérée en moyenne. Cependant, ils n'ont pas spécifiquement ciblé les utilisateurs de GS et n'ont pas proposé de solution formelle à ce problème. (Diggelen et al., 2005) A utilisé l'ensemble de données MovieLens pour tester cette approche pour un domaine CF en utilisant. S'agissant d'une simulation, ils n'ont pas décrit de méthode pour identifier ces utilisateurs et répondre à leurs besoins. Les utilisateurs de GS peuvent être reconnus à l'aide d'algorithmes de clustering hors ligne, où le seuil de similarité pour séparer ces utilisateurs du reste des clusters peut être trouvé empiriquement (Ghazanfar & Prugel-bennett, n.d.).

Pour identifier les utilisateurs de GS dans le système, de nombreuses approches sont suggérées, notamment la réutilisation de techniques de détection de valeurs aberrantes basées sur la distribution de similarité utilisateur-utilisateur (Zheng et al., 2017). Il s'agit d'une technique d'identification basée sur la distribution pour les utilisateurs de GS qui emprunte à la détection des valeurs aberrantes et à la recherche d'informations, tout en prenant en compte les spécificités des données de préférence sur lesquelles CF s'appuie (Gras et al., 2016), c'est-à-dire les approches basées sur le clustering (Ghazanfar & Prügel-Bennett, 2014)(Ghazanfar & Prugel-bennett, n.d.) ou les approches de réseaux sociaux (Minsung & Il, 2014). Dans l'ensemble, ces approches identifient les utilisateurs GS avec précision et éliminent les utilisateurs GS tout en faisant des recommandations pour le reste des utilisateurs avec un degré élevé de précision. Par conséquent, ils ne prennent pas en compte les utilisateurs de GS. Afin de résoudre ce problème, ce travail déploie tous les utilisateurs et bénéficie des utilisateurs de GS.

3. Notre Approche

Comme indiqué précédemment, l'approche CF de base utilise les K-voisins les plus proches pour faire de nouvelles prédictions. Il ne repose que sur les utilisateurs qui ont des préférences similaires à celles de l'utilisateur actif, quels que soient les utilisateurs présentant une faible similitude ou dissemblance dans la phase de prédiction. Dans les cas GS, la similitude entre l'utilisateur actif et les autres utilisateurs est faible ou inexistante, car la plupart d'entre eux sont éloignés. La figure 3 montre concrètement le cas du GS dans le processus CF.

Les trois cas possibles sont ainsi délimités :

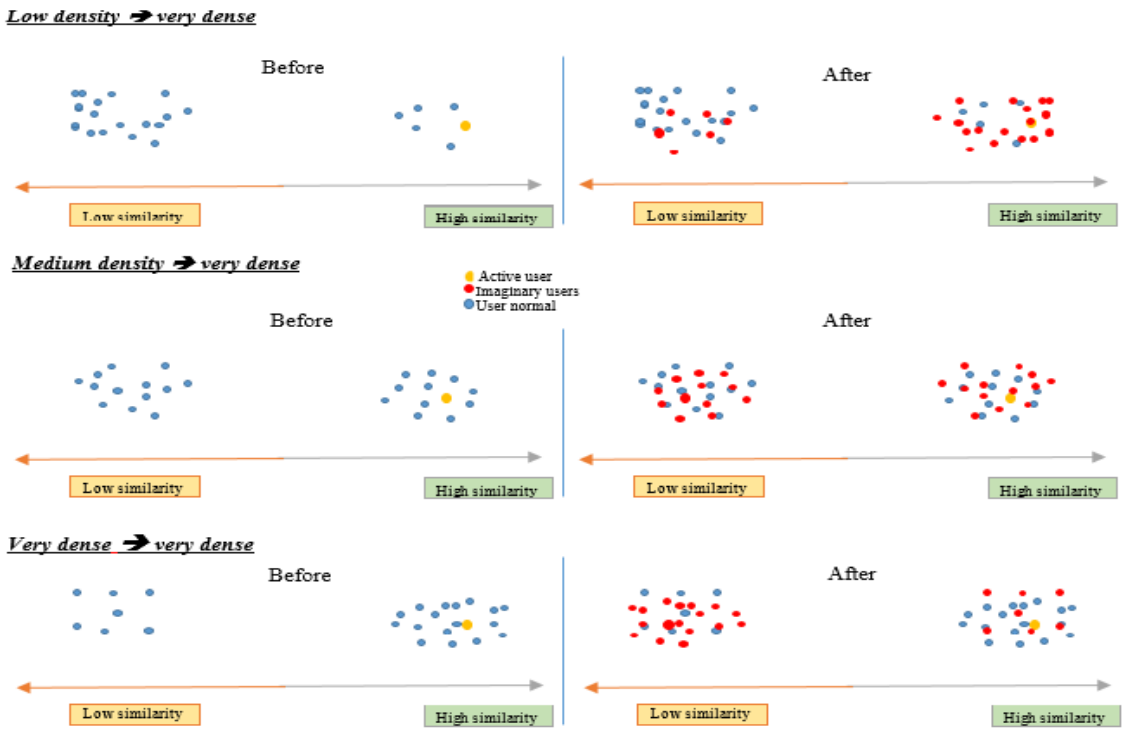


Figure 14: Exemple de cas de GS dans les techniques basées sur le voisinage

La figure ci-dessus montre un exemple de voisins hypothétiques qui peuvent être générés après avoir inversé les préférences. Nous avons représenté les nouveaux voisins imaginaires par des points rouges, formés via l'étape d'augmentation de la matrice. Les nouveaux voisins peuvent être positivement corrélés avec l'utilisateur cible.

Le cœur de notre approche est de profiter aux utilisateurs (utilisateurs GS) dont les préférences sont différentes de l'utilisateur cible. L'hypothèse sous-jacente de notre approche est que les utilisateurs doivent avoir plus ou moins les mêmes intérêts. Si les intérêts de l'utilisateur X sont opposés à ceux de l'utilisateur Y, alors l'utilisateur imaginaire ¬X aurait les mêmes intérêts que l'utilisateur Y. Par conséquent, des informations supplémentaires seront fournies au moteur de recommandation pour faire de bonnes recommandations.

Le nouveau processus de notre approche comprend une étape supplémentaire qui augmente le numéro de quartier de l'utilisateur actif qui est appelée augmentation de la matrice de notation (Figure 15).

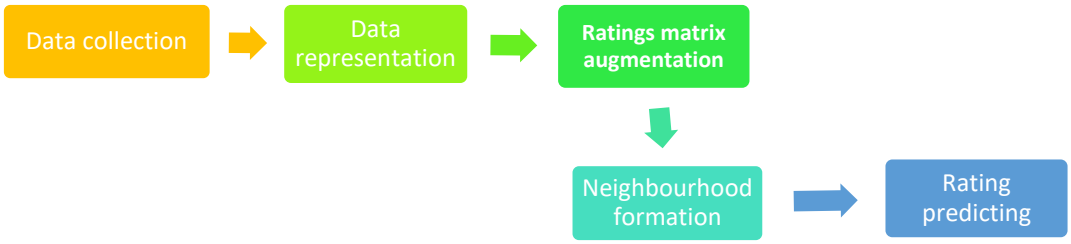


Figure 15: Nouveau processus CF basé sur la mémoire

L'étape d'augmentation de la matrice d'évaluation consiste à ajouter des lignes dans la matrice d'évaluation qui représentent les utilisateurs opposés aux utilisateurs réels. L'utilisateur imaginaire est obtenu en déduisant la préférence inverse de chaque élément évalué à l'aide de la formule suivante :

$$\neg r_{uj} = Max - r_{uj} + Min \quad (23)$$

r_{uj}: la note de l'utilisateur **u** pour un élément **j**.

Max et ***Min***: les valeurs haute et basse respectivement dans une échelle numérique donnée.

Example:

En fournissant un échantillon de notation allant de 1 à 5 dans le tableau ci-dessous, si un utilisateur **u** évalue un élément comme r_{uj}= 5 la note estimée de l'utilisateur **¬u** sera $\neg r_{uj}$=1.

Tableau 3: Exemple de matrice d'évaluation opposée sur une échelle de 5 points

Items / Users	I_1	I_2	I_3	I_4	I_5	I_6	I_7
U	**4**		**2**		**5**		**3**



Items / Users	I_1	I_2	I_3	I_4	I_5	I_6	I_7
U	**4**		**2**		**5**		**3**
¬*U*	**2**		**4**		**1**		**3**

Le tableau 3 illustre un exemple de cas de voisins imaginaires pouvant être générés après inversion des préférences. Nous avons représenté les nouveaux voisins imaginaires par des points rouges, formés via l'étape d'augmentation de la matrice. Les nouveaux voisins peuvent ainsi être positivement corrélés avec l'utilisateur cible. La figure 7 montre un exemple d'utilisateur opposé sur une échelle de 5 points utilisant la formule 6 ci-dessus. ¬U représente l'utilisateur opposé à l'utilisateur U après application de la formule. Il repose sur l'inférence des évaluations d'un utilisateur opposé en fournissant la préférence opposée d'un utilisateur donné.

Enfin, nous listons les pseudonymes de notre méthode proposée dans Algorithm Matrix Augmentation.

```
Algorithm : Matrix Augmentation
Input : Matrix T[nbrusers][nbritems] ;
Output: MatrixAugmentation TA [2 x nbrusers][nbritems];
Initialize Max, Min ;
cpt1 =1
cpt2 =1
while cpt 1 <=  nbrusers do // for all users in the dataset
     while cpt2  <=  nbritems do // for all his/her rated items
```

```
            if T[cpt1][ cpt2]>0
                  TA[nbrusers+cpt1]  [cpt2]=  Max  -  T[cpt1][  cpt2]
                  +Min
            cpt2++
      cpt1++
return TA
```

Figure 16: Algorithme d'augmentation de la matrice

4. Expérimentation et discussion

Dans cette section, de nombreuses expériences sont réalisées pour démontrer la nouveauté et l'efficacité de notre approche. Par conséquent, nous avons divisé notre ensemble de données en 80 % pour l'ensemble d'apprentissage et 20 % pour l'ensemble de test. Nous avons calculé les moyennes des résultats d'une validation croisée de 10 fois. Nous avons également implémenté un système de recommandation de films sous R grâce à Recommenderlab (Hahsler, 2011) avec les datasets MovieLens et FilmTrust. L'objectif est de vérifier les performances de notre approche proposée (AUBCF) avec l'approche CF traditionnelle basée sur l'utilisateur (UBCF) en utilisant des datasets du monde réel. Une brève description des datasets utilisés sera dans l'ordre, suivie de la procédure d'évaluation et de l'environnement de test de spécification. Par conséquent, les résultats ont été acquis à partir de comparaisons pour trouver l'approche la plus réussie.

4.1. Datasets

Nous avons exécuté nos expériences avec deux datasets couramment utilisés : FilmTrust et MovieLens. Les deux sont des projets de recherche académique de systèmes de recommandation de films basés sur le Web.

MovieLens est une dataset d'évaluation sur une échelle de 5 points allant de 1 (mauvais) à 5 (excellent). Il comprend 1682 films, 943 utilisateurs et 100 000 ratings.

La dataset FilmTrust comprend 1856 utilisateurs, 2092 films et 759922 ratings. Il a été collecté à partir d'un réseau social basé sur un système de recommandation vidéo comprenant des ratings. Les évaluations sont des échelles de 5 points allant de 0,5 à 4 étoiles.

4.2. Expériences

Toutes les expériences et techniques ont été réalisées sur Intel i5 à 2,4 GHz et 8 Go de RAM, en utilisant les datasets MovieLens et FilmTrust.

L'évaluation expérimentale de notre méthode suggérée est effectuée dans cette section, et les résultats sont basés sur une variété de mesures fréquemment utilisées avec divers paramètres.

4.2.1. Corrélation de Pearson :

La figure ci-dessous (Figure 17) montre les résultats obtenus en comparant notre approche proposée nommée AUBCF (UBCF augmenté) et l'approche CF basée sur l'utilisateur (UBCF) comme approche de base pour dataset FilmTrust.

La figure 17 représente une comparaison de MAE où l'axe horizontal est la taille du voisinage utilisé pour le calcul de MAE. La figure montre que notre approche (AUBCF) diminue régulièrement pour le MAE, tandis que l'approche traditionnelle (UBCF) diminue jusqu'à N = 40 puis reste stable jusqu'à N = 60 où le MAE commence à augmenter.

Sur la figure 18, nous pouvons voir que les MAE de notre approche (AUBCF), en vert, et de l'approche traditionnelle (UBCF), en rouge, sont inversement proportionnelles au nombre d'usagers dans le quartier. L'approche traditionnelle (UBCF) a un MAE plus élevé que notre approche (AUBCF).

Experiment 1: FilmTrust

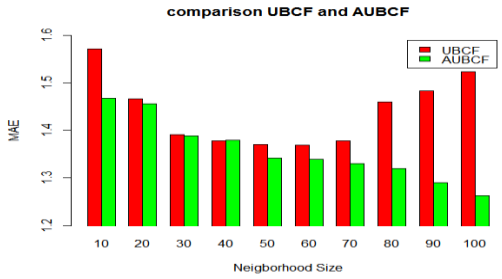


Figure 17: Comparaison de la MAE à l'aide de l'ensemble de données FilmTrust

Experiment 2: MovieLens

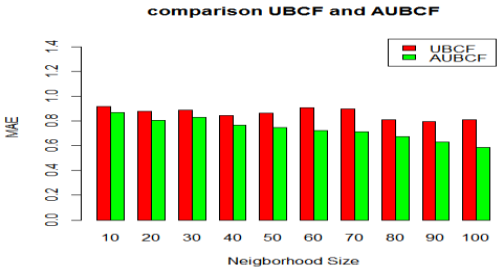


Figure 18: Comparaison de la MAE à l'aide de l'ensemble de données MovieLens

4.2.2. Cosinus:

Les figures 19 et 20 montrent les résultats obtenus en comparant l'approche traditionnelle (UBCF), en rouge, et l'approche proposée (AUBCF), en vert, pour chaque jeu de données. Le diagramme représente une comparaison de MAE où l'axe horizontal est la taille du voisinage dans chaque expérience. Il passe de 10 à 100 par intervalle de 10.

Experiment 1 : FilmTrust

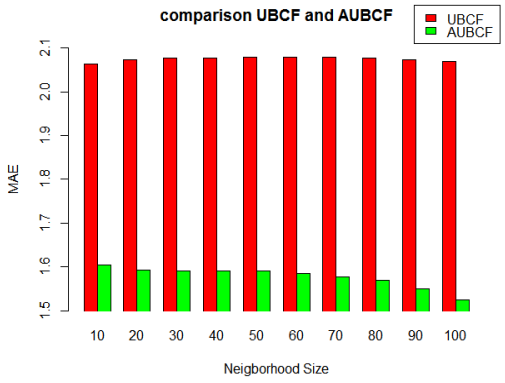


Figure 19: Comparaison de la MAE à l'aide de l'ensemble de données FilmTrust

Experiment 2 : MovieLens

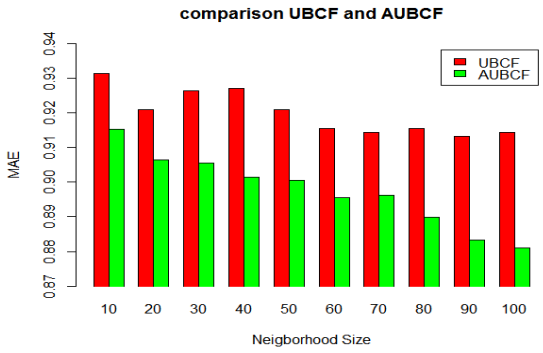


Figure 20: Comparaison de la MAE à l'aide de l'ensemble de données MovieLens

L'approche proposée (AUBCF) sur la figure 20 reste régulière pour le MAE alors que l'approche traditionnelle (UBCF) passe à N = 30 ; puis il reste stable jusqu'à N = 60 où le MAE commence à diminuer. Dans la figure 19, les MAE de l'approche proposée (AUBCF) et de l'approche traditionnelle (UBCF) sont inversement proportionnelles à la taille du quartier. Ils diminuent régulièrement jusqu'à N=60, puis ils restent stables jusqu'à N=100. Ainsi, notre approche (AUBCF) a une MAE plus faible que l'approche traditionnelle (UBCF).

Dans l'ensemble, nous concluons de ces expériences que l'approche proposée (AUBCF) offre de meilleures performances que l'approche traditionnelle (UBCF) dans les deux jeux de données.

4.3. Teste statistique

Dans la plupart des expériences, il est important de s'assurer que la différence observée entre la méthode proposée et la ligne de base est statistiquement significative ; et il est peu probable que cela soit dû au hasard ou au bruit dans les données. Le test statistique approprié à utiliser est le test de Wilcoxon. Le test de Wilcoxon est un test statistique non paramétrique qui compare deux échantillons de données sans supposer que les données ont une distribution spécifique, le but du test est de décider si les distributions de population sont identiques ou non. Notre hypothèse nulle est que les résultats de l'algorithme AUBCF et les résultats de l'algorithme UBCF sont des populations identiques, que tout petit gain, ou perte, observé n'est pas statistiquement significatif.

Généralement, on rejette l'hypothèse nulle si la p-value est inférieure à un certain seuil (souvent 0,05). En d'autres termes, si p-value < 0,05, nous pouvons en déduire que la

différence est statistiquement significative.

Tableau 4: Comparaison de la valeur p entre la corrélation de Pearson et cosinus par le test de Wilcoxon pour les deux ensembles de données

	Pearson correlation	Cosine
FilmTrust	6.900406e-25	0.01639486??????
MovieLens	4.473234e-05	0.0002432247

Selon le tableau, toutes les valeurs de p sont inférieures au seuil (0,05), nous avons rejeté l'hypothèse nulle et nous pouvons dire que la différence est statistiquement significative. Enfin, les résultats obtenus de l'algorithme AUBCF sont meilleurs que les résultats de l'algorithme UBCF.

5. Conclusion

Malgré la popularité et la grande utilisation du CF, il n'est pas sans limites car il doit encore surmonter le problème du GS. Par conséquent, ce travail a proposé une nouvelle approche CF pour résoudre ce problème. Cette approche vise à augmenter le nombre de voisins pour l'utilisateur actif en fonction des utilisateurs ayant des intérêts et des préférences différents. Pour évaluer notre algorithme, nous l'avons comparé à UBCF en tant qu'approche traditionnelle. La comparaison a été effectuée sur deux ensembles de données, FilmTrust et MovieLens. Les résultats obtenus montrent que notre approche surpasse UBCF et améliore la précision des prédictions pour les problèmes GS.

L'apport de notre travail peut se résumer en trois points principaux. Premièrement, notre approche utilise pleinement les données de notation pour améliorer la précision des systèmes de recommandation. Toutes les données d'évaluation des utilisateurs sont utilisées

dans le modèle, pas seulement les données d'évaluation des utilisateurs WS. Deuxièmement, le problème des utilisateurs GS est résolu dans notre modèle ce qui permet d'obtenir une similarité précise lorsqu'il n'y a pas de corrélation entre deux utilisateurs. Troisièmement, ce chapitre propose une nouvelle approche pour le filtrage collaboratif qui montre des performances supérieures au filtrage collaboratif traditionnel.

CHAPITRE 4 : FILTRAGE COLLABORATIF AMELIORE A BASE DE MODELE POUR L'ADRESSAGE DU PROBLEME DE MOUTON GRIS

Dans ce chapitre, nous avons opté, pour l'utilisation de l'approche basée sur un modèle. Notre concept fondamental réside dans la combinaison de l'inférence des utilisateurs imaginaires avec l'approche du filtrage collaboratif basée sur un modèle, ouvrant ainsi de nouvelles perspectives pour améliorer la qualité des recommandations.

1. Introduction

Au cours des dernières années, l'utilisation des systèmes de recommandation a connu une expansion significative, notamment dans les secteurs du commerce électronique et des médias en ligne. De nombreuses entreprises ont intégré ces systèmes pour optimiser leurs ventes en proposant des suggestions ou en recommandant des produits et services additionnels à leurs utilisateurs. En conséquence, ces systèmes ont révolutionné les interactions entre les entreprises en ligne et leur clientèle, marquant ainsi la transition des magasins physiques vers le monde numérique. Cette évolution a contraint de nombreuses entreprises à ajuster leurs stratégies pour répondre aux exigences des consommateurs (Ben Schafer et al., 2002).

Dans le domaine du commerce électronique, l'objectif principal des systèmes de recommandation est d'aider les clients à effectuer des choix d'achat éclairés tout en

simplifiant la longue liste de produits à une sélection plus personnalisée. Ces systèmes se manifestent sous forme de suggestions ou d'évaluations de produits sur des sites de vente en ligne tels qu'Amazon (Linden et al., 2017), Netflix (Gomez-uribe & Hunt, 2015), et Spotify (Celma, 2008)(Callan et al., 2003). Ces plateformes ont rencontré un succès significatif en rendant le divertissement et les achats plus accessibles aux consommateurs, offrant une expérience remarquable, particulièrement pendant la pandémie de Covid-19.

À notre connaissance, il existe trois principaux types des systèmes de recommandation: Les systèmes basés sur le contenu, les systèmes collaboratifs et les systèmes hybrides. Bien qu'elle soit efficace et simple, l'approche du filtrage collaboratif (CF) reste la plus utilisée dans les systèmes de recommandation (Ekstrand, 2011).

L'hypothèse principale qui sous-tend le filtrage collaboratif est basée sur des utilisateurs ou des éléments ayant des intérêts et des préférences similaires. Ces préférences peuvent être communiquées de différentes manières, notamment par des commentaires explicites (c'est-à-dire des évaluations, des appréciations, etc.) (Jawaheer et al., 2006), ou par des commentaires implicites déduits du comportement des utilisateurs (historique des achats, temps passé sur le contenu du site web, etc.) (Y. Hu et al., 2008). Ils peuvent être présentés sous la forme d'une matrice appelée matrice d'évaluation (Isinkaye et al., 2015b).

Deux types d'approches du filtrage collaboratif ont été proposés : le filtrage collaboratif basé sur la mémoire et le filtrage collaboratif basé sur un modèle. Dans le cas du filtrage collaboratif basé sur la mémoire, les recommandations sont générées sur la base de l'ensemble de la matrice d'évaluation (Paterek, 2007)(Agrawal & Agrawal, 2015)(Zheng et al., 2017). Par ailleurs, l'approche basée sur un modèle utilise des algorithmes

d'apprentissage automatique pour construire le modèle qui générera des éléments appropriés pour chaque utilisateur de manière personnalisée (P. Wei et al., 2019). Cette approche utilise des techniques d'apprentissage automatique telles que les règles d'association (AR), les techniques de regroupement (Breese et al., 1998), la factorisation des matrices (décomposition en valeurs singulières (SVD), décomposition en valeurs singulières de Funk (Funk SVD)), etc. Malgré les améliorations apportées aux méthodes de filtrage collaboratif, celles-ci rencontrent encore de nombreux défis qui limitent la qualité des recommandations, notamment les problèmes de mouton gris et de matrice creuse.

La matrice creuse fait référence au fait de disposer d'une grande quantité de données avec très peu d'évaluations ou d'interactions (Bobadilla et al., 2013), tandis que le mouton gris fait référence à une catégorie d'utilisateurs qui présentent des préférences ou des comportements qui s'écartent de manière significative de la population d'utilisateurs au sens large (Kant et al., 2018).

Dans ce travail, nous proposons une nouvelle approche qui consiste à combiner les méthodes SVD et les règles d’association avec une matrice d'évaluation enrichie.

Cet article se concentre sur les méthodes d'apprentissage automatique utilisées dans le filtrage collaboratif qui est largement utilisé aujourd'hui dans ce domaine : SVM, SVMF et AR. Enfin, nous présentons notre approche qui surpasse de nombreuses approches de recommandation traditionnelles et résout les problèmes de matrice creuse et de mouton gris.

Le reste de l'article est organisé comme suit : dans la section suivante, nous présentons les méthodes traditionnelles de CF, suivies d'une vue d'ensemble des travaux connexes

concernant les défis des moutons gris et matrice creuse dans les systèmes de recommandation dans la section 3. Dans la section 4, nous décrivons notre approche conçue pour relever les défis de la matrice creuse et du mouton gris. Ensuite, nous présentons notre analyse expérimentale, qui comprend des mesures d'évaluation et divers résultats expérimentaux dans la section 5. Enfin, dans la section 6, nous tirons des conclusions et suggérons des orientations potentielles pour la recherche future.

2. Travaux connexes

Dans cette section, nous abordons les questions de l'éparpillement et des utilisateurs de moutons gris, ce qui comble une lacune dans la recherche actuelle sur les systèmes de recommandation. Nous abordons également la question de l'obligation d'évaluer des utilisateurs opposés et la manière dont cette idée est exploitée dans la recherche sur les systèmes de recommandation. Nous mentionnons ensuite certains travaux qui examinent la CF à l'aide des techniques SVD et AR.

Selon (Claypool et al., 1999b), l'efficacité de l'algorithme CF traditionnel n'est pas uniforme d'un utilisateur à l'autre. Les utilisateurs peuvent être classés en deux groupes principaux : Les moutons blancs (WS) et les moutons gris (GS). Les utilisateurs WS présentent une grande similarité avec de nombreux autres utilisateurs, ce qui se traduit par un coefficient de corrélation élevé. En revanche, les moutons gris se caractérisent par leur dissemblance ou leur similarité partielle avec les autres utilisateurs, ce qui se traduit par des coefficients de corrélation plus faibles avec la plupart des utilisateurs (Adamopoulos, 2014). Par conséquent, les recommandations pour les utilisateurs GS ont tendance à être moins précises (Srivastava et al., 2020), et ces utilisateurs ne bénéficient donc pas pleinement des

systèmes de recommandation traditionnels. Bien que certains efforts aient été déployés pour résoudre le problème des utilisateurs GS (Ghorbani & Novin, 2016)(Ghazanfar & Prugel-bennett, n.d.), (Claypool et al., 1999b) ont mis l'accent sur le problème des utilisateurs GS et ont introduit un système de recommandation hybride visant à fournir des recommandations actualisées. Ce système hybride combine les approches de filtrage collaboratif (CF) et de filtrage basé sur le contenu par le biais d'une méthode de pondération moyenne. Toutefois, il convient de noter que cette approche ne cible pas spécifiquement les utilisateurs de GS et n'offre pas de solution formelle pour ce problème particulier (Srivastava, 2014). En outre, leur méthode a été testée à l'aide de l'ensemble de données MovieLens dans un domaine CF, mais il s'agissait principalement d'une simulation. Ils n'ont pas décrit de méthode pour identifier les utilisateurs de SG ou pour répondre à leurs besoins particuliers. L'utilisation d'algorithmes de regroupement hors ligne constitue une approche potentielle de l'identification des utilisateurs du SG. En employant des méthodes de regroupement, il est possible d'identifier les utilisateurs GS sur la base de seuils de similarité empiriques, en les séparant efficacement des autres groupes d'utilisateurs (Gong et al., 2009). Le problème de la matrice creuse se pose lorsque la matrice d'évaluation, qui contient des détails essentiels sur les évaluations, est particulièrement peu dense. Cette rareté peut entraîner des inefficacités dans les systèmes de recommandation. Ce problème se divise en deux catégories : la couverture limitée et la transitivité (Prabhu et al., 2018). La matrice creuse est un problème qui concerne les RS en général et les CF en particulier. Pour surmonter ce problème, diverses approches et méthodes ont été proposées. La majorité d'entre elles reçoivent un retour d'information positif. Toutefois, il n'existe toujours pas de solution générale pour traiter la matrice creuse (Guo et al., 2017)(Natarajan et al., 2020)(Hwangbo & Kim, 2017)(N. Mishra et al., 2017).

En outre, les algorithmes de règles d'association (AR) permettent d'établir des relations entre les éléments d'une transaction. Ces règles, telles que A --> B, sont applicables dans les modèles de recommandation pour anticiper les caractéristiques de l'utilisateur et de l'article (Guo et al., 2017)(Natarajan et al., 2020). La stratégie d'exploration de données basée sur les règles d'association est celle qui est la plus étudiée (S. Zhang & Wu, 2011). Les règles d'association spécifient la manière dont un événement est lié à un autre. Il s'agit également d'une sorte de regroupement qui catégorise les données en fonction de leur importance (Jooa et al., 2019). Lorsque l'événement A se produit, une règle d'association est donnée sous la forme R : A -> B, indiquant que l'événement B peut également se produire. Cette combinaison peut réduire la taille et le temps de traitement des données massives.

Dans le contexte des systèmes de recommandation, la décomposition en valeurs singulières (SVD) est une méthode bien établie utilisée pour relever les défis du filtrage collaboratif (Vozalis & Margaritis, 2007). La SVD est une approche populaire du filtrage de recommandation qui décompose une matrice m x n A en trois matrices à l'aide de la formule $A = USV^T$ (Vozalis & Margaritis, 2007).

La SVD de Funk, une variante de la SVD régularisée, vise à créer une décomposition matricielle qui se rapproche étroitement des valeurs de la matrice réelle avec une erreur minimale (M. Yu et al., 2022). Il convient de noter que la Funk SVD a permis de remporter le concours de recommandation Netflix (Simo Funk, 2006). La Funk SVD est une forme spécialisée de la SVD conçue pour les systèmes de recommandation. Elle intègre des techniques d'optimisation et est conçue pour traiter les problèmes spécifiques associés aux

données de recommandation, ce qui la rend plus efficace dans ce contexte. Simon Funk (Simo Funk, 2006) a proposé une SVD régularisée appelée Funk SVD pour le filtrage collaboratif, une méthodologie inspirée de méthodes efficaces du domaine du traitement du langage naturel (Linden et al., 2017). La proposition de taux d'apprentissage et de constantes de régularisation, ainsi qu'une méthode d'écrêtage des prédictions, sont incluses dans la description de Simon Funk (H. Chen et al., 2021). Le modèle de décomposition de la valeur singulière de Funk (F-SVD) est une variante de la factorisation de la matrice (MF) introduite par Funk en 2006. Il s'est fait connaître par ses performances supérieures lors du concours Netflix Prize, qui visait à améliorer les algorithmes de recommandation de films. Le concept central de la F-SVD, comme dans la MF traditionnelle, est de représenter les utilisateurs et les éléments à l'aide de vecteurs de caractéristiques latents dérivés de la matrice d'évaluation. Lorsqu'il existe une forte correspondance entre les vecteurs de caractéristiques des utilisateurs et des éléments, il est plus facile de formuler des recommandations précises (H. Chen et al., 2021).

L'application de méthodes de réduction des dimensions à la matrice d'évaluation est la méthodologie la plus répandue (Zhu et al., 2016)(Ghosh & Chakraborty, 2017). Le problème de la rareté est traité par la méthode de réduction des dimensions, qui condense la matrice en supprimant les données peu représentatives ou bruyantes.

D'autres méthodes, telles que l'indexation sémantique latente (LSI), est fréquemment utilisée (Kumar et al., 2016). L'un des inconvénients de ces approches est qu'au cours du processus de réduction, des informations potentiellement utiles peuvent être inévitablement supprimées. L'utilisation de techniques de recherche associative vers le graphe bipartite

composé d'éléments et d'utilisateurs (X. Hu et al., 2017), la CF boostée par le contenu (Gautam et al., 2016), les techniques de décomposition en valeurs singulières (SVD) (Sahoo et al., 2019) et le levage de la similarité basée sur les éléments pour remplacer la similarité basée sur les utilisateurs (Shambour, 2016) sont des méthodes supplémentaires pour traiter la rareté.

Ces méthodes construisent un modèle qui apprend des interactions entre l'utilisateur et l'élément en utilisant des représentations à faible dimension comme composante (vecteurs de caractéristiques de l'utilisateur et de l'élément) (Sharma et al., 2017). Le modèle SVD inclut des données d'utilisateurs imaginaires pour améliorer la matrice d'évaluation dans le processus de factorisation de la matrice, ce qui permet d'obtenir de meilleures recommandations que le modèle SVD traditionnel.

Les travaux de (El Fazziki et al., 2022) et (El Fazziki et al., 2020) confirment que l'efficacité de l'algorithme CF traditionnel est inférieure à celle de l'approche basée sur l'augmentation de la matrice d'évaluation. Cette dernière est réalisée par l'utilisation d'utilisateurs imaginaires, qui sont l'opposé des utilisateurs réels. La même idée est utilisée dans (El Fazziki et al., 2019) et confirme que cette idée améliore la prédiction et résout le problème des moutons gris. Cependant, les deux travaux utilisent l'approche CF basée sur la mémoire.

Plusieurs lacunes techniques générales observées dans les travaux existants ont conduit à la conception de l'approche proposée. Tout d'abord, les systèmes de recommandation disponibles ne se concentrent pas sur l'efficacité ou la réalisation des objectifs ultimes pour chaque système de recommandation, en particulier pour la

recommandation de films.

Néanmoins, il est impératif de noter que les approches mentionnées ci-dessus ne fournissent pas de solution précise aux problèmes de matrice creuse et d'utilisateurs de moutons gris, et tendent à ignorer ces problèmes. Par conséquent, en présence de problèmes tels que la matrice creuse et les utilisateurs "moutons gris", les systèmes de recommandation ne trouvent pas suffisamment de données pour fournir des recommandations efficaces. C'est pourquoi, dans cette étude, nous présentons notre approche pour résoudre ces problèmes.

3. Notre approche

Comme indiqué précédemment, l'approche CF de base (approche traditionnelle) dépend uniquement de la base de données d'évaluation, que ce soit pour prédire la similarité entre l'utilisateur et l'élément ou pour former un modèle. Étant donné que la plupart des éléments ne sont évalués que par quelques utilisateurs - l'une des raisons de la rareté et de la grisaille des utilisateurs et de la matrice utilisateur-élément qui en résulte - est relativement due à un manque de données d'évaluation, ce qui rend les algorithmes inefficaces.

Nos recherches s'appuient sur les fondements posés par (El Fazziki et al., 2022). L'idée fondamentale qui sous-tend notre approche tourne autour de l'incorporation d'une phase supplémentaire dans le processus traditionnel de filtrage collaboratif, que nous appelons "matrice d'évaluation enrichie". Il convient de noter que les travaux présentés dans (El Fazziki et al., 2022) sont principalement axés sur une approche basée sur la mémoire. En revanche, notre étude adopte une approche basée sur le modèle, en tirant parti de techniques

telles que l'AR, la SVD et la FSVD pour faire progresser et élargir le concept d'enrichissement de la matrice d'évaluation.

Notre approche révisée introduit une phase supplémentaire appelée "matrice d'évaluation enrichie" (El Fazziki et al., 2022) ((Figure 21), qui vise à étendre la matrice d'évaluation.



Figure 21: Nouveau processus de l'approche du CF

La figure 21 illustre le nouveau processus proposé par l'approche CF avec l'ajout d'une nouvelle phase appelée matrice d'évaluation enrichie.

Notre méthode enrichie est fondée sur le concept selon lequel les consommateurs doivent partager des intérêts similaires. Si les intérêts de l'utilisateur X sont totalement contraires à ceux de l'utilisateur Y, l'utilisateur imaginaire ¬X aurait ces intérêts. Par conséquent, davantage de données seront envoyées au moteur de recommandation afin de suggérer des recommandations potentielles et appropriées.

L'étape d'enrichissement de la matrice d'évaluation consiste à ajouter des lignes à la matrice qui représentent des utilisateurs opposés aux utilisateurs réels. L'utilisateur imaginaire est créé en calculant la préférence opposée de chaque élément évalué à l'aide de la formule :

$$\neg r_{uj} = Max - r_{uj} + Min \qquad (24)$$

r_{uj}: la note de l'utilisateur **u** pour un item **j**.

Max et ***Min***: les valeurs hautes et basses respectivement dans une échelle numérique donnée.

Exemple:

Si un utilisateur u note un article avec une note de r_{uj}= 5, la note anticipée de l'utilisateur u sera $\neg r_{uj}$= 1.. Si un utilisateur u note un article avec une note de r_{uj}= 1., la note anticipée de l'utilisateur u sera $\neg r_{uj}$= 5.

Tableau 5: Exemple de matrice d'évaluation opposée sur une échelle de 5 points

Items / Users	I_1	I_2	I_3	I_4	I_5	I_6	I_7
U	4		2		5		3
$\neg U$	2		4		1		3

Nous améliorons la densité de la matrice d'évaluation en incorporant les préférences opposées des utilisateurs. Cette matrice enrichie est ensuite utilisée en combinaison avec les techniques de SVD et de règles d'association pour résoudre efficacement les problèmes liés à la matrice creuse et à la présence de moutons gris parmi les utilisateurs.

4. Résultats expérimentaux

Dans cette section, nous décrivons la mesure des performances de l'ensemble de données et les résultats de la comparaison entre l'approche CF que nous proposons et les approches CF traditionnelles.

Dans cette section, plusieurs tests ont été effectués pour démontrer l'originalité et l'efficacité de notre approche. En conséquence, nous divisons notre ensemble de données

en deux parties : 80 % pour la formation et 20 % pour le test (figure 22) ; ce concept est connu sous le nom de "Train/Test Split".

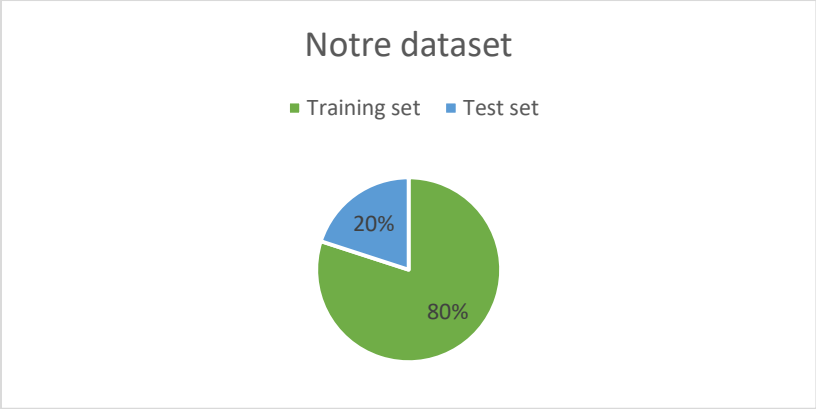


Figure 22: Dataset divisées

En utilisant le package Recommenderlab (Zhu et al., 2016) et le jeu de données MovieLens, nous avons développé avec succès un système de recommandation de films en R.

L'objectif est de comparer les performances de l'approche que nous proposons à celles des approches traditionnelles en utilisant des ensembles de données du monde réel. L'approche d'évaluation et l'environnement de test spécifique seront suivis d'une brève discussion sur les ensembles de données utilisés. Par conséquent, les résultats ont été dérivés de comparaisons afin de déterminer l'approche la plus efficace.

4.1. Collecte d'ensembles de données :

L'ensemble de données MovieLens est utilisé pour évaluer les performances de la technique proposée.

Le projet de recherche GroupLens de l'université du Minnesota (Adomavicius & Tuzhilin, 2005) a généré les ensembles de données MovieLens, et MovieLens est un

système de recommandation de recherche basé sur le web qui a été lancé à l'automne 1997; les données sont accessibles gratuitement à l'adresse suivante : https://grouplens.org/datasets/movielens/.

Chaque semaine, des centaines de personnes visitent MovieLens pour évaluer et recevoir des recommandations de films. Les évaluations se font sur une échelle de cinq points, 1 et 2 indiquant les évaluations négatives, 4 et 5 les évaluations positives et 3 l'ambivalence [27]. Il comprend 1682 films, 943 utilisateurs et 100 000 classements.

4.2. Expériences

L'évaluation expérimentale de la méthode que nous proposons est réalisée dans cette section, et les résultats sont basés sur une variété de métriques fréquemment utilisées avec différents paramètres.

Tous les algorithmes ont été écrits dans le langage de programmation R et exécutés sur un processeur Intel i7 2,4 GHz avec 8 Go de RAM sous Windows 8.1, en utilisant les datasets MovieLens. Le choix de R pour construire le système de recommandation de films est justifié par ses capacités statistiques robustes, ses bibliothèques d'apprentissage automatique spécialisées comme RecommenderLab, et son soutien important au sein de la communauté de la science des données, fournissant un environnement efficace pour le développement et le prototypage d'algorithmes de recommandation.

4.3. Résultats et discution

Nous appliquons des mesures de précision statistique dans le présent document (Gong et al., 2009). Les mesures de précision statistique comparent l'écart numérique entre les

évaluations prédites et les évaluations réelles des utilisateurs afin de déterminer la précision d'un algorithme de prédiction. Dans le contexte de l'évaluation des systèmes de recommandations, des mesures telles que l'erreur absolue moyenne (MAE), l'erreur quadratique moyenne (RMSE) et l'erreur quadratique moyenne (MSE) sont fréquemment employées. Ces mesures permettent d'évaluer quantitativement l'exactitude et la précision d'un système de recommandation. La MAE calcule les différences absolues moyennes entre les évaluations prédites et réelles, la RMSE mesure la racine carrée des différences moyennes au carré, et la MSE représente les différences moyennes au carré entre les évaluations prédites et réelles. Ces mesures jouent un rôle crucial dans l'évaluation des performances et de l'efficacité des algorithmes de recommandation. Toutes les mesures susmentionnées ont été calculées à l'aide des mêmes données (ensembles de données MovieLens) et ont produit des résultats similaires.

Les mesures MAE, MSE et RMSE calculent les différences entre nos résultats et cette réalité facétieuse.

L'objectif de cette section est d'évaluer la MAE, la MSE et la RMSE pour chaque ensemble d'observations.

Tableau 6: Comparaison des RMSE, MSE et MAE entre AR et enriched_AR pour 10 expériences

Testing set		RMSE	MSE	MAE
Fold 1	AR	1,852	2,032	1,342
	Enriched_AR	1,502	1,82	1,025
Fold 2	AR	1,822	2,056	1,369
	Enriched_AR	1,523	1,798	1,009
Fold 3	AR	1,789	2,112	1,379
	Enriched_AR	1,498	1,852	1,011
Fold 4	AR	1,955	2,089	1,396
	Enriched_AR	1,598	1,85	1,023

Fold 5	AR	2,002	2,189	1,401
	Enriched_AR	1,6	1,935	1,156
Fold 6	AR	1,785	2,006	1,299
	Enriched_AR	1,598	1,796	1,123
Fold 7	AR	1,986	2,103	1,388
	Enriched_AR	1,536	1,758	1,099
Fold 8	AR	1,896	2,156	1,379
	Enriched_AR	1,489	1,832	1,008
Fold 9	AR	1,763	2,098	1,399
	Enriched_AR	1,423	1,864	1,096
Fold 10	AR	1,754	2,179	1,389
	Enriched_AR	1,453	1,897	1,097
Average	***AR***	***1,8604***	***2,102***	***1,3741***
	Enriched_AR	***1,522***	***1,8402***	***1,0647***

Le tableau 6 présente une analyse comparative des résultats de 10 expériences réalisées à l'aide de règles d'association et de règles d'association enrichies. Les mesures d'évaluation utilisées pour la comparaison sont RMSE, MSE et MAE.

Les résultats montrent que les valeurs RMSE, MSE et MAE obtenues avec l'approche des règles d'association enrichies sont systématiquement inférieures à celles obtenues avec les règles d'association. Cette valeur inférieure pour les trois mesures d'évaluation conduit collectivement à une réduction des résultats moyens pour la RA enrichie par rapport à ceux de la AR. Cela indique que l'approche de la AR enrichie est plus performante que l'approche de la AR standard en termes de précision prédictive. La figure ci-dessous illustre les résultats moyens de dix expériences, fournissant une comparaison entre la AR enrichie et la AR traditionnelle.

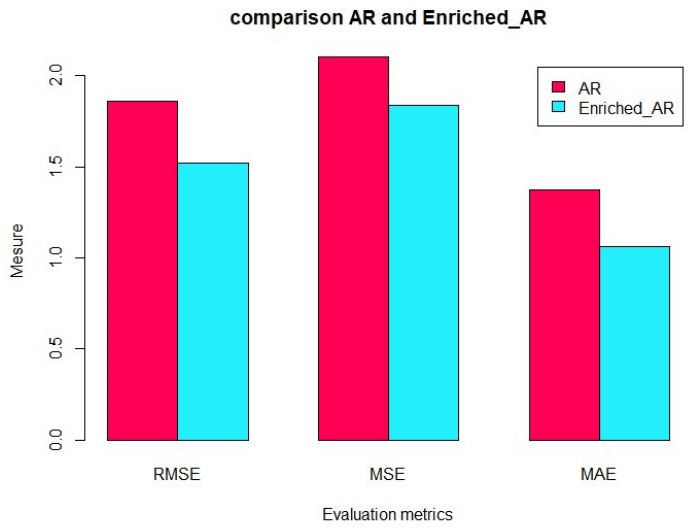


Figure 23:Comparaison entre AR et Enriched_AR

La figure 23 présente une comparaison détaillée entre l'approche que nous avons proposée, la règle d'association enrichie (enriched_AR), et la méthode traditionnelle, la règle d'association (AR). Le graphique à barres fournit une représentation visuelle de la comparaison entre ces deux approches sur la base des mesures d'évaluation moyennes RMSE, MSE et MAE.

Dans le premier scénario, représenté à la figure 23, les valeurs RMSE montrent que la méthode AR obtient un score de 1,86, tandis que la méthode enriched_AR est plus performante, avec une valeur inférieure de 1,52. Dans le deuxième cas, si l'on se concentre sur l'EQM, AR atteint 2,1, tandis que enriched_AR obtient un score inférieur de 1,84. Enfin, dans le troisième cas, en examinant la métrique MAE, AR obtient 1,37, tandis qu'enriched_AR excelle avec un score nettement inférieur de 1,06. Ces résultats indiquent de manière cohérente que en-riched_AR est plus performant que AR dans les trois scénarios.

La représentation graphique de ces résultats dans le diagramme à barres montre clairement que, pour toutes les mesures d'évaluation (RMSE, MSE et MAE), AR produit systématiquement des valeurs supérieures à celles de enriched_AR. Cela prouve clairement que l'approche que nous proposons, enriched_AR, est supérieure à l'approche traditionnelle, AR.

Tableau 7: Comparaison des RMSE, MSE et MAE entre SVD et enriched_SVD pour 10 expériences

Testing set		RMSE	MSE	MAE
Fold 1	SVD	1,21011	1,25655	0,89236
	Enriched_SVD	0,7435	0,94162	0,47278
Fold 2	SVD	1,24478	1,2775	0,94858
	Enriched_SVD	0,74024	0,98497	0,49057
Fold 3	SVD	1,21982	1,23805	0,9204
	Enriched_SVD	0,73189	0,92665	0,44087
Fold 4	SVD	1,20832	1,23142	0,95088
	Enriched_SVD	0,74439	0,96595	0,45655
Fold 5	SVD	1,20244	1,25993	0,90302
	Enriched_SVD	0,71236	0,98328	0,41388
Fold 6	SVD	1,2325	1,23727	0,9634
	Enriched_SVD	0,72077	0,93343	0,40325
Fold 7	SVD	1,20086	1,2997	0,94676
	Enriched_SVD	0,71125	0,95438	0,4095
Fold 8	SVD	1,22357	1,28151	0,93541
	Enriched_SVD	0,74187	0,95739	0,46279
Fold 9	SVD	1,21788	1,2579	0,9028
	Enriched_SVD	0,72557	0,97042	0,43356
Fold 10	SVD	1,23759	1,2579	0,94779
	Enriched_SVD	0,70121	0,99907	0,42491
Average	***SVD***	***1,219787***	***1,259773***	***0,93114***
	Enriched_SVD	***0,727305***	***0,961716***	***0,440866***

Le tableau 7 présente une analyse complète des résultats obtenus à partir de 10 expériences comparant la SVD et la SVD enrichie. Les mesures d'évaluation utilisées dans cette analyse sont la RMSE, la MSE et la MAE.

Les résultats révèlent que les valeurs de RMSE, MSE et MAE obtenues à l'aide de l'approche SVD enrichie sont systématiquement inférieures à celles obtenues à l'aide de la SVD standard. Cette réduction constante des valeurs pour les trois mesures d'évaluation contribue à une diminution globale des résultats moyens de l'approche SVD enrichie par rapport à l'approche SVD standard. Cela signifie que l'approche SVD enrichie surpasse l'approche SVD conventionnelle en termes de précision prédictive. La figure suivante illustre les résultats moyens de dix expériences, présentant une comparaison entre l'approche SVD enrichie et l'approche SVD traditionnelle.

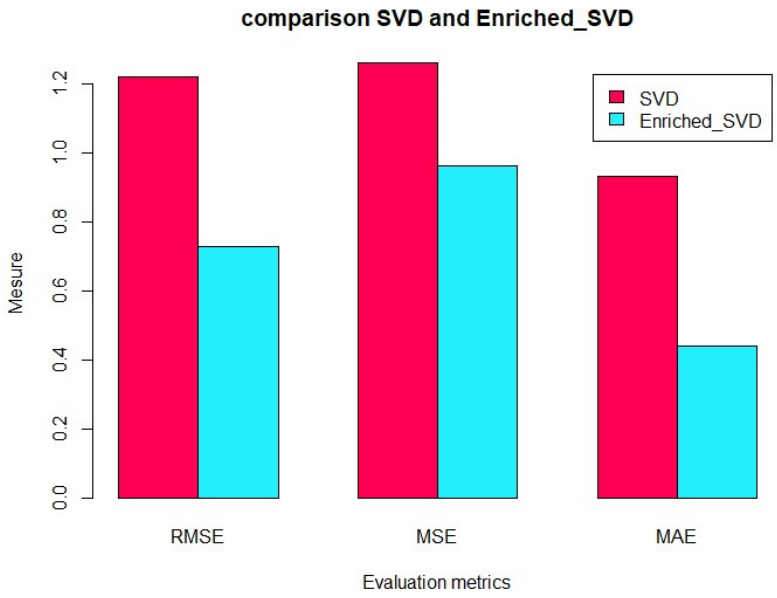


Figure 24: Comparaison entre SVD et Enriched_SVD

Dans la figure 24, nous présentons une analyse comparative des mesures moyennes RMSE, MSE et MAE entre nos deux approches : SVD représentée par les barres rouges, et en-riched_SVD représentée par les barres bleues. L'axe x du graphique présente les mesures d'évaluation RMSE, MSE et MAE, tandis que l'axe y représente une échelle de mesure allant de 0,0 à 2,0.

Pour la RMSE, la barre rouge, correspondant à la SVD, indique une valeur de 1,21 sur l'échelle de mesure moyenne le long de l'axe des abscisses. En revanche, la barre bleue, qui représente SVD Enrichie, affiche une valeur nettement inférieure de 0,72, ce qui indique une performance supérieure à celle de la SVD en termes de RMSE.

De même, lors de l'évaluation du MSE, la barre rouge pour la SVD affiche une valeur de 1,25 sur l'axe des abscisses. En revanche, la barre bleue pour la SVD enrichie affiche une valeur inférieure de 0,96, ce qui met encore plus en évidence l'amélioration des performances de la SVD enrichie par rapport à la méthode SVD en ce qui concerne le MSE.

Enfin, en ce qui concerne la MAE, la barre rouge, représentant la SVD, indique une valeur de 0,93 sur l'axe des abscisses. En revanche, la barre bleue, représentant la SVD enrichie, révèle une valeur considérablement réduite de 0,44, ce qui souligne l'amélioration des performances de la SVD enrichie par rapport à la méthode SVD en ce qui concerne l'MAE. Les résultats de ce graphique montrent donc que notre approche (enriched_SVD) est plus efficace et plus performante que l'approche traditionnelle (SVD) dans tous les cas (RMSE, MSE et MAE).

Tableau 8: Comparaison des RMSE, MSE et MAE entre FSVD et enriched_FSVD pour 10 expériences

Testing set		RMSE	MSE	MAE
fold1	FSVD	1,0995458	1,2090009	0,8685648
	Enriched_FSVD	0,94073	0,94762	0,8656485
fold 2	FSVD	1,085267	1,187536	1,084247
	Enriched_FSVD	0,91999	0,81466	0,64318
fold 3	FSVD	1,104534	1,18234	0,90273
	Enriched_FSVD	0,82788	0,89096	0,70782
fold 4	FSVD	1,1000983	1,07594	0,99912
	Enriched_FSVD	0,87149	0,99031	0,74489

fold 5	FSVD	1,1056543	1,12792	0,94007
	Enriched_FSVD	0,95691	0,93999	0,67979
fold 6	FSVD	1,1007852	1,02487	0,86891
	Enriched_FSVD	0,81719	0,84763	0,60187
fold 7	FSVD	1,0827826	1,14485	0,94529
	Enriched_FSVD	0,87548	0,99797	0,67487
fold 8	FSVD	1,1209832	1,13386	0,84648
	Enriched_FSVD	0,95354	0,95324	0,71967
fold 9	FSVD	1,084165	1,09896	0,91404
	Enriched_FSVD	0,86949	0,86621	0,69407
fold 10	FSVD	1,0651261	1,10651	0,82199
	Enriched_FSVD	0,83651	0,88177	0,62448
Average	**FSVD**	**1,09489415**	**1,12917869**	**0,91914418**
	Enriched_FSVD	**0,886921**	**0,913036**	**0,69562885**

Le tableau 8 présente une analyse approfondie des résultats de 10 expériences comparant le FSVD et le FSVD enrichi. Les résultats montrent systématiquement que les valeurs de RMSE, MSE et MAE obtenues avec l'approche FSVD enrichie sont nettement inférieures à celles obtenues avec l'approche FSVD standard. Cette réduction constante des valeurs pour les trois paramètres d'évaluation contribue à une diminution globale des résultats moyens pour l'approche FSVD enrichie par rapport à l'approche FSVD standard. Cela suggère que l'approche FSVD enrichie est supérieure à l'approche FSVD conventionnelle en termes de précision prédictive.

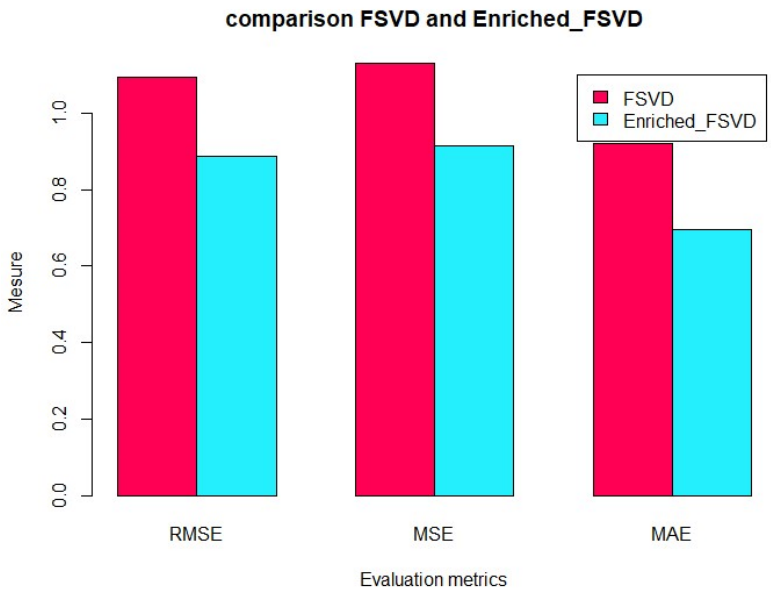


Figure 25: Comparaison entre FSVD et Enriched_FSVD

La figure 25 présente une comparaison visuelle des mesures d'évaluation moyennes RMSE, MSE et MAE pour FSVD et enrichi_FSVD. Ces mesures sont représentées sur l'axe des x, tandis que l'échelle de mesure se trouve sur l'axe des y. Dans chacun des trois cas (RMSE, MSE et MAE), la barre représentant FSVD est systématiquement plus haute que la barre pour Enriched_FSVD. Cette constance dans les trois mesures permet de conclure que notre approche, enrichie_FSVD, est plus performante que l'approche traditionnelle (FSVD) en toutes circonstances.

Toutes les figures montrent que notre approche a une MAE, une RMSE et une MSE inférieures à l'approche traditionnelle pour l'ensemble de données MovieLens.

4.4. Inférences statistiques

Dans la plupart des expériences, il est primordial de s'assurer que la différence observée entre la méthode proposée et la référence est statistiquement significative. Cela

permet de confirmer que la disparité n'est pas simplement le résultat du hasard ou d'un bruit aléatoire dans les données. Le test statistique le plus approprié pour de telles comparaisons est le test t (test de Student). Le test t est une méthode statistique non paramétrique, c'est-à-dire qu'il ne repose pas sur des hypothèses spécifiques de répartition des données. Son objectif principal est de déterminer si les distributions de la population sont identiques ou différentes. Dans ce contexte, notre hypothèse nulle postule que les résultats obtenus par l'algorithme enrichi et l'algorithme traditionnel proviennent de populations identiques, ce qui exclut tout gain ou perte minime dû au hasard.

Notre hypothèse nulle postule que les résultats dérivés des algorithmes enrichis et ceux des algorithmes traditionnels appartiennent à des populations identiques, ce qui signifie que les améliorations ou détériorations marginales observées sont statistiquement insignifiantes. Selon la pratique habituelle, nous rejetons l'hypothèse nulle lorsque la valeur p est inférieure à un seuil prédéfini, souvent fixé à 0,05. En termes plus simples, si la valeur p est inférieure à 0,05, nous pouvons raisonnablement conclure que les disparités sont statistiquement significatives.

Nous avons donc comparé les valeurs p entre RMSE, MSE et MAE à l'aide du test t, et les résultats sont présentés dans le tableau 9.

Tableau 9: Comparaison de la valeur p entre RMSE, MSE et MAE par t -test pour les méthodes enrichis et les méthodes traditionnels

	RMSE	MSE	MAE
Enriched-AR/AR	**8,11096455E-08**	**8,39879E-09**	**4,63195E-08**
Enriched-SVD/ SVD	**4,2122E-14**	**3,31869E-11**	**1,54672E-11**
Enriched-FSVD/ FSVD	**1,23187E-07**	**6,71672E-06**	**2,88269E-06**

Comme l'indique le tableau 9, toutes les valeurs p calculées sont inférieures au seuil

de 0,05. Par conséquent, nous rejetons l'hypothèse nulle dans tous les cas, ce qui prouve de manière irréfutable que les différences observées sont effectivement statistiquement significatives.

5. Conclusion

Cette approche enrichie vise à augmenter la matrice d'évaluation en fonction des intérêts et des préférences des utilisateurs. Pour évaluer nos méthodes, nous les avons comparées à l'approche traditionnelle en utilisant l'ensemble de données MovieLens. Notre approche est plus performante que les approches traditionnelles et améliore la précision de la prédiction. Cela confirme que la méthode enrichie donne de meilleurs résultats que la méthode traditionnelle. Les résultats de l'étude actuelle montrent que l'innovation dans les systèmes de recommandation peut être obtenue par l'efficacité et la combinaison d'approches nouvelles et anciennes, comme le révèle la tendance en termes de recherche.

Nos recherches en cours se concentrent sur l'exploration des approches d'apprentissage automatique en ligne, un domaine émergent qui a le potentiel de fournir des informations essentielles sur la dynamique des bases de consommateurs en évolution. Dans nos travaux futurs, nous prévoyons d'approfondir les techniques supplémentaires de détection des valeurs aberrantes afin d'examiner plus en détail l'évolution des profils d'utilisateurs au fil du temps.

CONCLUSION ET TRAVAUX FUTURE

En plus de la conclusion de cette étude, ce chapitre propose des recommandations pour les futures recherches dans le domaine de cette thèse et se structure comme suit. En premier lieu, la section 1 synthétise les conclusions de la thèse et récapitule les contributions apportées, en mettant en relief la distinction de nos approches par rapport aux travaux existants sur les systèmes de recommandation. Ensuite, la section 2 expose les orientations futures de la recherche.

En conclusion, les systèmes de recommandation sont devenus une composante essentielle de nombreuses plateformes et services en ligne. Ces systèmes visent à aider les utilisateurs à découvrir des éléments pertinents, tels que des films, des produits, de la musique, et plus encore, en fonction de leurs préférences et de leurs intérêts.

Les systèmes de recommandation utilisent divers algorithmes et techniques, notamment le filtrage collaboratif, le filtrage basé sur le contenu et autres. Ils exploitent les données de l'utilisateur, telles que son comportement passé, ses évaluations et ses informations démographiques, ainsi que les attributs des articles, les informations contextuelles et les connexions sociales, afin de générer des recommandations personnalisées.

Les motivations qui sous-tendent le développement et l'utilisation des systèmes de recommandation sont multiples. Elles comprennent l'amélioration de la

satisfaction et de l'engagement des utilisateurs, l'augmentation des ventes et des revenus, l'amélioration de l'expérience de l'utilisateur et la découverte efficace et personnalisée d'articles. En suggérant des articles qui correspondent aux préférences des utilisateurs, les systèmes de recommandation visent à faire gagner du temps aux utilisateurs et à leur épargner des efforts dans la recherche d'articles pertinents, ce qui se traduit par une expérience utilisateur plus agréable et plus personnalisée.

Cependant, les systèmes de recommandation sont également confrontés à des défis et à des limites telles que la matrice creuse, les problèmes de démarrage à froid, les moutons gris et la possibilité de créer des bulles de filtre ou de renforcer les préjugés existants sont des sujets de préoccupation. Les considérations éthiques, la protection de la vie privée et le besoin de transparence dans les algorithmes de recommandation sont également des sujets d'intérêt dans ce domaine.

Néanmoins, les systèmes de recommandation continuent d'évoluer et de s'améliorer, en intégrant les progrès de l'apprentissage automatique, du traitement du langage naturel et de l'analyse des données massives (big data). L'intégration de diverses sources de données, telles que le comportement des utilisateurs, les caractéristiques des articles, les connexions sociales et les informations contextuelles, a permis de formuler des recommandations plus précises, plus pertinentes et plus personnalisées.

Les systèmes de recommandation jouent un rôle crucial en aidant les utilisateurs à naviguer dans l'énorme quantité d'informations disponibles et à trouver des articles

qui correspondent à leurs préférences. Grâce à la recherche et au développement continus, les systèmes de recommandation ont le potentiel d'améliorer encore les expériences des utilisateurs, de stimuler l'engagement et de faciliter la découverte personnalisée dans divers domaines.

Dans le cadre de nos perspectives qui guideront nos futurs travaux dans ce domaine, nous envisageons sérieusement l'intégration des techniques de Deep Learning dans les approches du filtrage collaboratif. Cette démarche vise à explorer les potentialités offertes par le Deep Learning pour améliorer la précision et la robustesse de nos systèmes de recommandation.

Une autre perspective prometteuse que nous souhaitons explorer est l'investigation des techniques de GNN (Graph Neural Networks) en synergie avec notre idée de base. Cette approche innovante pourrait permettre une représentation plus riche des relations entre utilisateurs et éléments dans le réseau, ouvrant ainsi de nouvelles voies pour renforcer l'efficacité du filtrage collaboratif.

Notre engagement actuel se manifeste dans un article en cours de développement, où nous intégrons spécifiquement les GNN dans le contexte du filtrage collaboratif. Nous sommes convaincus que cette fusion de technologies ouvrira des perspectives nouvelles et substantielles dans le domaine des systèmes de recommandation, et nous sommes impatients de partager les résultats de cette recherche avec la communauté académique.

REFERENCES

Adamopoulos, P. (2014). On discovering non-obvious recommendations: Using unexpectedness and neighborhood selection methods in collaborative filtering systems. *WSDM 2014 - Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, 655–659. https://doi.org/10.1145/2556195.2556204

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. https://doi.org/10.1109/TKDE.2005.99

Adomavicius, G., & Tuzhilin, A. (2015). Context-aware recommender systems. *Recommender Systems Handbook, Second Edition*, 191–226. https://doi.org/10.1007/978-1-4899-7637-6_6

Agrawal, S., & Agrawal, J. (2015). Survey on anomaly detection using data mining techniques. *Procedia Computer Science*, *60*(1), 708–713. https://doi.org/10.1016/j.procs.2015.08.220

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval. *New York*, *9*, 513. https://doi.org/10.1080/14735789709366603

Balabanovi, M., & Shoham, Y. (1997). Content-based, Collaborative Recommendation. *Communications of the ACM*, *40*(3), 66--72.

Basile, P., De Gemmis, M., Gentile, A. L., Iaquinta, L., & Lops, P. (2008). The JUMP project: Domain ontologies and linguistic knowledge @ work. *CEUR Workshop Proceedings*, *314*(May 2014).

Beel, J., & Langer, S. (2015). A comparison of offline evaluations, online evaluations, and user studies in the context of research-paper recommender systems. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9316*, 153–

168. https://doi.org/10.1007/978-3-319-24592-8_12

Behera, D. K., Das, M., Swetanisha, S., & Sethy, P. K. (2021). Hybrid model for movie recommendation system using content K-nearest neighbors and restricted boltzmann machine. *Indonesian Journal of Electrical Engineering and Computer Science*, *23*(1), 445–452. https://doi.org/10.11591/ijeecs.v23.i1.pp445-452

Belkin, N., & Croft, W. (1992). Information filtering and information retrieval: two sides of the same coin? *Communications of the ACM*, *29*(10), 1–10. https://doi.org/10.1145/138859.138861

Bhaidani, S. (2008). Recommender system algorithms. *Proceedings of the International Conference on Weblogs and Social Media ICWSM 2007*. http://www.mie.utoronto.ca/undergrad/thesis-catalog/files/146.pdf

Billsus, D., & Pazzani, M. (1997). Learning Probabilistic User Models. *Proceedings of the Workshop on Machine Learning for User Modeling*. citeseer.nj.nec.com/billsus96learning.html

Billsus, D., & Pazzani, M. J. (1999). A hybrid user model for news story classification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *407*, 99–108. https://doi.org/10.1007/978-3-7091-2490-1_10

Billsus Daniel, & Pazzani Michael J. (2002). User modeling for adaptative news access. . *User Modeling and User-Adapted Interaction*, *10*, 147–180. www.fxpal.com/people/billsus/pubs/um4news.pdf

Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. (2013). Recommender systems survey. *Knowledge-Based Systems*, *46*, 109–132. https://doi.org/10.1016/j.knosys.2013.03.012

Breese, J., Heckerman, D., & Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence*, 43--52. https://doi.org/10.1111/j.1553-2712.2011.01172.x

Burke, R. (2002). Hybrid Recommender Systems : Survey and. *User Modeling and UserAdapted Interaction*, *12*(4), 331–370. https://doi.org/10.1023/A:1021240730564]

Burke, R. (2010). Evaluating the dynamic properties of recommendation algorithms. *RecSys'10 - Proceedings of the 4th ACM Conference on Recommender Systems*, 225–228. https://doi.org/10.1145/1864708.1864753

Burke, R., & Hurley, N. (2011). Robust collaborative recommendation. In *Recommender Systems Handbook* (Issue March 2014). https://doi.org/10.1007/978-0-387-85820-3

Callan, J., Smeaton, A., Beaulieu, M., Borlund, P., Brusilovsky, P., Chalmers, M., Lynch, C., Riedl, J., Smyth, B., Straccia, U., & Toms, E. (2003). Personalisation and Recommender Systems in Digital Libraries Joint NSF-EU DELOS Working Group Report. *Library*, *5*(May), 299–308. https://doi.org/10.1007/s00799-004-0100-1

Canny, J. (2002). Collaborative Filtering with Privacy via Factor Analysis Categories and Subject Descriptors. *IEEE Security & Privacy*.

Cantador, I., Bellogín, A., & Castells, P. (2008). News@hand: A semantic web approach to recommending news. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5149 LNCS*(July), 279–283. https://doi.org/10.1007/978-3-540-70987-9_34

Celma, O. (2008). Music recommendation and discovery in the long tail. *Citeulikeorg*, 252. http://www.citeulike.org/user/bemike/article/4099800

Chaudhari, K., & Thakkar, A. (2020). A Comprehensive Survey on Travel Recommender Systems. *Archives of Computational Methods in Engineering*, *27*(5), 1545–1571. https://doi.org/10.1007/s11831-019-09363-7

Chen, H., Sheng, S., Xu, C. Y., Li, Z., Zhang, W., Wang, S., & Guo, S. (2021). A spatiotemporal

estimation method for hourly rainfall based on F-SVD in the recommender system. *Environmental Modelling and Software*, *144*, 105148. https://doi.org/10.1016/j.envsoft.2021.105148

Chen, R., Hua, Q., Chang, Y. S., Wang, B., Zhang, L., & Kong, X. (2018). A survey of collaborative filtering-based recommender systems: from traditional methods to hybrid methods based on social networks. *IEEE Access*, *6*, 64301–64320. https://doi.org/10.1109/ACCESS.2018.2877208

Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., & Sartin, M. (1999a). Combining content-based and collaborative filters in an online newspaper. *Proceedings of the ACM SIGIR '99 Workshop on Recommender Systems: Algorithms and Evaluation*.

Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., & Sartin, M. (1999b). Combining content-based and collaborative filters in an online newspaper. *Proceedings of the ACM SIGIR '99 Workshop on Recommender Systems: Algorithms and Evaluation*, *June*.

Degemmis, M., Lops, P., & Semeraro, G. (2007). A content-collaborative recommender that exploits WordNet-based user profiles for neighborhood formation. In *User Modeling and User-Adapted Interaction* (Vol. 17, Issue 3). https://doi.org/10.1007/s11257-006-9023-4

Diggelen, J. Van, Dignum, V., Elst, L. Van, & Abecker, A. (2005). *Proceedings of the AAMAS 2005 Workshop on Agent Mediated Knowledge Management*. *July*.

Eirinaki, M., Vazirgiannis, M., & Varlamis, I. (2003). SEWeP: Using site semantics and a taxonomy to enhance the Web personalization process. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *January*, 99–108. https://doi.org/10.1145/956750.956765

Ekstrand, M. D. (2011). Collaborative Filtering Recommender Systems. *Foundations and Trends® in Human–Computer Interaction*, *4*(2), 81–173. https://doi.org/10.1561/1100000009

El Fazziki, A., El Madani El Alami, Y., El Aissaoui, O., El Allioui, Y., & Benbrahim, M. (2020).

Improving Collaborative Filtering Approach by Leveraging Opposite Users. In *Advances in Intelligent Systems and Computing: Vol. 1102 AISC* (Issue January). Springer International Publishing. https://doi.org/10.1007/978-3-030-36653-7_14

El Madani El Alami, Y., Nfaoui, E. H., & Beqqali, O. El. (2016). A novel collaborative filtering approach based on social network experts. *2016 International Conference on Information Technology for Organizations Development, IT4OD 2016*, 1–7. https://doi.org/10.1109/IT4OD.2016.7479277

El Madani El Alami, Y., Nfaoui, E. H., & El Beqqali, O. (2015). Toward an effective hybrid collaborative filtering: A new approach based on matrix factorization and heuristic-based neighborhood. *2015 Intelligent Systems and Computer Vision, ISCV 2015*. https://doi.org/10.1109/ISACV.2015.7105543

Fazziki, A. El, Aissaoui, O. El, Alami, Y. E. M. El, Allioui, Y. El, & Benbrahim, M. (2019). A new collaborative approach to solve the gray-sheep users problem in recommender systems. *2019 3rd International Conference on Intelligent Computing in Data Sciences, ICDS 2019*, 1–4. https://doi.org/10.1109/ICDS47004.2019.8942256

Fazziki, A. El, Alami, Y. E. M. El, Elhassouni, J., Aissaoui, O. El, & Benbrahim, M. (2022). Employing opposite ratings users in a new approach to collaborative filtering. *Indonesian Journal of Electrical Engineering and Computer Science*, *25*(1), 450–459. https://doi.org/10.11591/ijeecs.v25.i1.pp450-459

Gabrilovich, E., & Markovitch, S. (2006). Computing Semantic Relatedness Using Wikipedia based Explicit Semantic Analysis. *Proceedings of the National Conference on Artificial Intelligence*, *2*, 1301–1306.

Gautam, A., Chaudhary, P., Sindhwani, K., & Bedi, P. (2016). CBCARS: Content boosted context-aware recommendations using tensor factorization. *2016 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2016*, 75–81.

https://doi.org/10.1109/ICACCI.2016.7732028

Ghazanfar, M. A., & Prugel-bennett, A. (n.d.). *Fulfilling the Needs of Gray-Sheep Users in Recommender Systems , A Clustering Solution*.

Ghazanfar, M. A., & Prügel-Bennett, A. (2014). Leveraging clustering approaches to solve the gray-sheep users problem in recommender systems. *Expert Systems with Applications*, *41*(7), 3261–3275. https://doi.org/10.1016/j.eswa.2013.11.010

Ghorbani, S., & Novin, A. H. (2016). An Introduction on Separating Gray-Sheep Users in Personalized Recommender Systems Using Clustering Solution. *International Journal of Computer Science and Software Engineering*, *5*(2), 14–18.

Ghosh, A. K., & Chakraborty, A. (2017). Use of EM algorithm for data reduction under sparsity assumption. *Computational Statistics*, *32*(2), 387–407. https://doi.org/10.1007/s00180-016-0657-3

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, *35*(12), 61–70. https://doi.org/10.1145/138859.138867

Gomez-uribe, C. A., & Hunt, N. (2015). *The Netflix Recommender System_Algorithms, Business Value.pdf*. *6*(4). https://doi.org/10.1145/2843948

Gong, S. J., Wu Ye, H., & Tan, H. S. (2009). Combining memory-based and model-based collaborative filtering in recommender system. *Proceedings of the 2009 Pacific-Asia Conference on Circuits, Communications and System, PACCS 2009*, 690–693. https://doi.org/10.1109/PACCS.2009.66

Gras, B., Brun, A., & Boyer, A. (2016). *Identifying Grey Sheep Users in Collaborative Filtering*. 17–26. https://doi.org/10.1145/2930238.2930242

Gunes, I., Bilge, A., & Polat, H. (2013). Shilling attacks against memory-based privacy-preserving

recommendation algorithms. *KSII Transactions on Internet and Information Systems*, *7*(5), 1272–1290. https://doi.org/10.3837/tiis.2013.05.019

Guo, G., Qiu, H., Tan, Z., Liu, Y., Ma, J., & Wang, X. (2017). Resolving data sparsity by multi-type auxiliary implicit feedback for recommender systems. *Knowledge-Based Systems*, *138*, 202–207. https://doi.org/10.1016/j.knosys.2017.10.005

Hahsler, M. (2011). recommenderlab: A Framework for Developing and Testing Recommendation Algorithms. *Nov*, *December 2013*, 1–37. http://cran.r-project.org/web/packages/recommenderlab/vignettes/recommenderlab.pdf

Harper, F. M., & Konstan, J. A. (2015). The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, *5*(4). https://doi.org/10.1145/2827872

Hernando, A., Bobadilla, J., & Ortega, F. (2016). A non negative matrix factorization for collaborative filtering recommender systems based on a Bayesian probabilistic model. *Knowledge-Based Systems*, *97*, 188–202. https://doi.org/10.1016/j.knosys.2015.12.018

Hill, W., Stead, L., Rosenstein, M., & Furnas, G. (1995). Recommending and evaluating choices in a virtual community of use. *Conference on Human Factors in Computing Systems - Proceedings*, *1*, 194–201. https://doi.org/10.1145/223904.223929

Hu, X., Mai, Z., Zhang, H., Xue, Y., Zhou, W., & Chen, X. (2017). A hybrid recommendation model based on weighted bipartite graph and collaborative filtering. *Proceedings - 2016 IEEE/WIC/ACM International Conference on Web Intelligence Workshops, WIW 2016*, 119–122. https://doi.org/10.1109/WIW.2016.4

Hu, Y., Koren, Y., & Volinsky, C. (2008). Collaborative Filtering for Implicit Feedback Datasets. *Gastroenterology*, *1)*, S415. https://doi.org/10.1109/ICDM.2008.22

Hwangbo, H., & Kim, Y. (2017). An empirical study on the effect of data sparsity and data overlap on

cross domain collaborative filtering performance. *Expert Systems with Applications*, *89*, 254–265. https://doi.org/10.1016/j.eswa.2017.07.041

Iaquinta, L., Gentile, A. L., Lops, P., de Gemmis, M., & Semeraro, G. (2008). *A Hybrid Content-Collaborative Recommender System Integrated into an Electronic Performance Support System*. 47–52. https://doi.org/10.1109/his.2007.30

Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015a). Recommendation systems: Principles, methods and evaluation. *Egyptian Informatics Journal*, *16*(3), 261–273. https://doi.org/10.1016/j.eij.2015.06.005

Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015b). Recommendation systems: Principles, methods and evaluation. *Egyptian Informatics Journal*, *16*(3), 261–273. https://doi.org/10.1016/j.eij.2015.06.005

J. Ben Schafer, Joseph Konstan, and J. R. (1999). Recommender systems in E-Commerce. *Encyclopedia of Human Computer Interaction*. https://doi.org/10.4018/978-1-59140-562-7.ch073

Jawaheer, G., Szomszor, M., & Kostkova, P. (2006). *Comparison of implicit and explicit feedback from an online music recommendation service*. 47–51. https://doi.org/10.1145/1869446.1869453

Jindal, H., Agarwal, S., & Sardana, N. (2018). PowKMeans: A Hybrid Approach for Gray Sheep Users Detection and Their Recommendations. *International Journal of Information Technology and Web Engineering*, *13*(2), 56–69. https://doi.org/10.4018/IJITWE.2018040106

Jonathan L. Herlocker, Joseph A. Konstan, Al Borchers, and J. R. (1999). An algorithmic framework for performing collaborative filtering. *In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Vol. 51. 230–237.

Jooa, J., Bangb, S., & Parka, G. (2016). Implementation of a Recommendation System Using Association Rules and Collaborative Filtering. *Procedia Computer Science*, *91*(Itqm 2016), 944–952.

https://doi.org/10.1016/j.procs.2016.07.115

Jooa, J., Bangb, S., & Parka, G. (2019). Implementation of a Recommendation System using Association Rules and Collaborative Filtering. *Procedia - Procedia Computer Science*, *91*(Itqm 2016), 944–952. https://doi.org/10.1016/j.procs.2016.07.115

Kant, S., Mahara, T., Kumar Jain, V., Kumar Jain, D., & Kumar Sangaiah, A. (2018). LeaderRank based k-means clustering initialization method for collaborative filtering. *Computers and Electrical Engineering*, *69*, 598–609. https://doi.org/10.1016/j.compeleceng.2017.12.001

Kasinathan, V., Mustapha, A., Tong, T. S., Rani, M. F. C. A., & Rahman, N. A. A. (2019). Heartbeats: Music recommendation system with fuzzy inference engine. *Indonesian Journal of Electrical Engineering and Computer Science*, *16*(1), 275–282. https://doi.org/10.11591/ijeecs.v16.i1.pp275-282

Kobayashi, S. (2005). Ontological User Profiling in Recommender Systems. *ChemInform*, *36*(12). https://doi.org/10.1002/chin.200512266

Konstan, J. A., & Riedl, J. (2012). Recommender systems: From algorithms to user experience. *User Modeling and User-Adapted Interaction*, *22*(1–2), 101–123. https://doi.org/10.1007/s11257-011-9112-x

Koren, Y. (2008). *Factorization meets the neighborhood*. 426–434. https://doi.org/10.1145/1401890.1401944

Koren, Y. (2010). Timesvd++. *Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, *2052*, 447-455,.

Kumar, B., Srivastava, A., & Kumar, P. (2016). Cosine Based Latent Factor Model for Precision Oriented Recommendation. *International Journal of Advanced Computer Science and Applications*, *7*(1), 451–457. https://doi.org/10.14569/ijacsa.2016.070161

Landgrebe, T. C., Paclik, P., Duin, R. P., & Bradley, A. P. (2006). Precision-recall operating characteristic (P-ROC) curves in imprecise environments. In Proceedings of 18th International Conference on Pattern Recognition. Los Alamitos, California USA. *IEEE Computer Society*, *4*(c).

Lees-Miller, J., Anderson, F., Hoehn, B., & Greiner, R. (2008). Does Wikipedia information help Netflix predictions? *Proceedings - 7th International Conference on Machine Learning and Applications, ICMLA 2008*, 337–343. https://doi.org/10.1109/ICMLA.2008.121

Lewis, D. D. (1994). A Comparison of Two Learning Algorithms for Text Categorization 1 Introduction 2 Text Categorization : Nature and Approaches. *In Proceeding of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR`94)*, 1–14.

Li, Y., Huang, J., Jiang, S. H., Huang, F., & Chang, Z. (2017). Evaluating collaborative filtering recommender systems. *Scientific Reports*, *7*(1), 5–53. https://doi.org/10.1038/s41598-017-17507-7

Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, *7*(1), 76–80. https://doi.org/10.1109/MIC.2003.1167344

Linden, G., Smith, B., & York, J. (2017). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, *7*(1), 76–80. https://doi.org/10.1109/MIC.2003.1167344

Lops, P., Gemmis, M. De, & Semeraro, G. (2011). Content-based Recommender Systems: State of the Art and Trends. In *Recommender Systems Handbook*. https://doi.org/10.1007/978-0-387-85820-3

Lv, Y., & Zhai, C. X. (2009). Adaptive relevance feedback in information retrieval. *International Conference on Information and Knowledge Management, Proceedings*, *3*, 255–264. https://doi.org/10.1145/1645953.1645988

Maertens, R. M., Long, A. S., & White, P. A. (2017). A Comparison of Event Models for Na¨ıve Bayes Text Classification. *Environmental and Molecular Mutagenesis*, *58*(8), 582–591. https://doi.org/10.1002/em.22125

Magnini, B., & Strapparava, C. (2001). Improving user modelling with content-based techniques. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *2109*, 74–83. https://doi.org/10.1007/3-540-44566-8_8

Maidin, S. S., Fan, L. I. U., Lau, S., & Yew, B. (2024). *FROM THEORY TO PRACTICE : UNDERSTANDING THE FACTORS AFFECTING THE DEVELOPMENT OF DIGITAL COMMUNITY EDUCATION IN CHINA* . *102*(2), 386–397.

McNee, S. M., Riedl, J., & Konstan, J. A. (2006). Being accurate is not enough: How accuracy metrics have hurt recommender systems. *Conference on Human Factors in Computing Systems - Proceedings*, 1097–1101. https://doi.org/10.1145/1125451.1125659

Middleton, S. E., Shadbolt, N. R., & De Roure, D. C. (2004). Ontological user profiling in recommender systems. *ACM Transactions on Information Systems*, *22*(1), 54–88. https://doi.org/10.1145/963770.963773

Mihalcea, R., & Csomai, A. (2007). Wikify! Linking documents to encyclopedic knowledge. *International Conference on Information and Knowledge Management, Proceedings*, 233–241. https://doi.org/10.1145/1321440.1321475

Minsung, K., & Il, I. (2014). *Resolving the 'Gray sheep' Problem Using Social Network Analysis (SNA) in Collaborative Filtering (CF) Recommender Systems*. *20*(2), 137–148.

Mishra, B. K., Sahoo, A. K., & Misra, R. (2018). Recommendation for selecting smart village in India through opinion mining using big data analytics. *Advances in Intelligent Systems and Computing*, *653*, 105–112. https://doi.org/10.1007/978-981-10-6602-3_11

Mishra, N., Chaturvedi, S., Mishra, V., Srivastava, R., & Bargah, P. (2017). Solving sparsity problem in rating-based movie recommendation system. *Advances in Intelligent Systems and Computing*, *556*, 111–117. https://doi.org/10.1007/978-981-10-3874-7_11

Mobasher, B., Burke, R., Bhaumik, R., & Williams, C. (2005). Effective Attack Models for Shilling Item-Based Collaborative Filtering System. *Webkdd*, *June 2014*, 13–23. https://www.researchgate.net/publication/243787659

Montaner, M., López, B., & De La Rosa, J. L. (2003). A taxonomy of recommender agents on the internet. *Artificial Intelligence Review*, *19*(4), 285–330. https://doi.org/10.1023/A:1022850703159

Mooney, R. J., & Roy, L. (2000). Content-based book recommending using learning for text categorization. *Proceedings of the ACM International Conference on Digital Libraries*, *June*, 195–204. https://doi.org/10.1145/336597.336662

Natarajan, S., Vairavasundaram, S., Natarajan, S., & Gandomi, A. H. (2020). Resolving data sparsity and cold start problem in collaborative filtering recommender system using Linked Open Data. *Expert Systems with Applications*, *149*. https://doi.org/10.1016/j.eswa.2020.113248

Ngoupeyou Tondji, L. (2018). *Web Recommender System for Job Seeking and Recruiting*. *February*. https://doi.org/10.13140/RG.2.2.26177.61286

Ni, J., Cai, Y., Tang, G., & Xie, Y. (2021). Collaborative filtering recommendation algorithm based on TF-IDF and user characteristics. *Applied Sciences (Switzerland)*, *11*(20). https://doi.org/10.3390/app11209554

Ortega, F., Zhu, B., Bobadilla, J., & Hernando, A. (2018). Knowle dge-Base d Systems CF4J : Collaborative filtering for Java. *Knowledge-Based Systems*, *0*, 1–6. https://doi.org/10.1016/j.knosys.2018.04.008

Panniello, U., Tuzhilin, A., & Gorgoglione, M. (2014). Comparing context-aware recommender systems in terms of accuracy and diversity. *User Modeling and User-Adapted Interaction*, *24*(1–2), 35–65. https://doi.org/10.1007/s11257-012-9135-y

Paterek, A. (2007). Improving regularized singular value decomposition for collaborative filtering.

Proceedings of KDD Cup and Workshop, 2–5. https://doi.org/10.1145/1557019.1557072

Paul Resnick, Neophytos Lacovou, Mitesh Suchak, Peter Bergstrom, J. R. (1999). GroupLens: An Open Architecture for Collaborative Filtering of Netnews. *Research Papers in Economics (1994*. https://doi.org/10.2957/kanzo.58.183

Pazzani, M. J., & Billsus, D. (2007). Content-Based Recommendation Systems. *Construction and Building Materials*, *171*, 546–557. https://doi.org/10.1016/j.conbuildmat.2018.03.149

Pazzani, M., Muramatsu, J., & Billsus, D. (1996). Syskill & Webert: identifying interesting web sites. *Proceedings of the National Conference on Artificial Intelligence*, *1*, 54–61.

Pelikan, M., Sastry, K., & Goldberg, D. E. (2002). Scalability of the Bayesian optimization algorithm. *International Journal of Approximate Reasoning*, *31*(3), 221–258. https://doi.org/10.1016/S0888-613X(02)00095-6

Polatidis, N., & Georgiadis, C. K. (2017). A dynamic multi-level collaborative filtering method for improved recommendations. *Computer Standards and Interfaces*, *51*, 14–21. https://doi.org/10.1016/j.csi.2016.10.014

Prabhu, R., Shetty, P., Shwetha, D. R., & Hegde, R. (2018). *A review : Recommender System using Collaborative Filtering and Gray Sheep Problem*. *6*(2), 440–443.

Rao, K. N., & Talwar, V. G. (2008). Application Domain and Functional Classification of Recommender Systems — A Survey. *DESIDOC Journal of Library Information Technology*, *28*(3), 17–35. https://doi.org/10.14429/djlit.28.3.174

Resnick, P., & Varian, H. R. (1997). Recommender Systems Guest Editors. *Ecom Syst COMMUNICATIONS OF THE ACM*, *56*(3), 56–58.

Ricci, F., Rokach, L., & Shapira, B. (2011). Recommender Systems Handbook. In *Recommender Systems*

Handbook. https://doi.org/10.1007/978-0-387-85820-3

Ricci, F., Rokach, L., & Shapira, B. (2015). *Recommender Systems Handbook*. https://doi.org/10.1007/978-1-4899-7637-6

Riedl, J. (n.d.). *Fine tuning the social Web*. https://cse.umn.edu/college/feature-stories/fine-tuning-social-web-john-riedl

Sahoo, A. K., Pradhan, C., & Mishra, B. S. P. (2019). SVD based privacy preserving recommendation model using optimized hybrid item-based collaborative filtering. *Proceedings of the 2019 IEEE International Conference on Communication and Signal Processing, ICCSP 2019*, 294–298. https://doi.org/10.1109/ICCSP.2019.8697950

Salakhutdinov, R., Mnih, A., & Hinton, G. (2007). Restricted Boltzmann Machines for Collaborative Filtering. *Journal of Computing and Information Technology*, *6*(3), 245–254.

Salter, J., & Antonopoulos, N. (2006). CinemaScreen recommander agent: Combining collaborative and content-based filtering. *IEEE Intelligent Systems*, *21*(1), 35–41. https://doi.org/10.1109/MIS.2006.4

Salton, G. (1992). The state of retrieval system evaluation. *Information Processing and Management*, *28*(4), 441–449. https://doi.org/10.1016/0306-4573(92)90002-H

Samah, K. A. F. A., Sabri, N., Hamzah, R., Roslan, R., Mangshor, N. A., & Asri, A. A. M. (2019). Brute force algorithm implementation for traveljoy travelling recommendation system. *Indonesian Journal of Electrical Engineering and Computer Science*, *16*(2), 1042–1049. https://doi.org/10.11591/ijeecs.v16.i2.pp1042-1049

Sarwar B, Karypis G, Konstan J, R. J. (2001). Item-based collaborative filtering recommendation algorithms. *Proceedings of the 10th International Conference OnWorldWideWeb*, 285–295.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2000). Application of Dimensionality Reduction in

Recommender System A Case Study Technical Report CS-TR 00-043. *Computer Science and Engineering Dept., University of Minnesota*.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001a). Item-based collaborative filtering recommendation algorithms. *Proceedings of the 10th …*, *1*, 285–295. https://doi.org/10.1145/371920.372071

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001b). Item-based collaborative filtering recommendation algorithms. *Proceedings of the 10th International Conference on World Wide Web, WWW 2001*, 285–295. https://doi.org/10.1145/371920.372071

Science, C., & Wnek, J. (1997). Learning and Revising User Profiles: The Identification of Interesting Web Sites. *Machine Learning*, *331*, 313–331.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, *34*(1), 1–47. https://doi.org/10.1145/505282.505283

Semeraro, G., Lops, P., Basile, P., & De Gemmis, M. (2009). Knowledge infusion into content-based recommender systems. *RecSys'09 - Proceedings of the 3rd ACM Conference on Recommender Systems*, 301–304. https://doi.org/10.1145/1639714.1639773

Shambour, Q. (2016). *A User-Based Multi-Criteria Recommendation Approach for Personalized Recommendations*. *14*(12), 657–663.

Shardanand, U., & Maes, P. (1995). Social information filtering: algorithms for automating "word of mouth." *Proceedings of the SIGCHI Conference on …*, 1–15. http://dl.acm.org/citation.cfm?id=223931

Sharma, R., Gopalani, D., & Meena, Y. (2017). *Collaborative Filtering – Based Recommender System : Approaches and Research Challenges*. 1–6.

Simo Funk. (2006). *Netflix update: Try this at home*. Https://Sifter.Org/Simon/Journal/20061211.Html.

Smirnov, A. V., & Krizhanovsky, A. A. (2008). Information filtering based on wiki index database. *World Scientific Proceedings Series on Computer Engineering and Information Science 1; Computational Intelligence in Decision and Control - Proceedings of the 8th International FLINS Conference*, 115–120. https://doi.org/10.1142/9789812799470_0019

Smyth, B. (2007). Case-based recommendation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4321 LNCS*, 342–376. https://doi.org/10.1007/978-3-540-72079-9_11

Srivastava, A. (2014). Gray sheep, influential users, user modeling and recommender system adoption by startups. *J Intell Inform Syst*, *20*(no.2), 137–148. https://doi.org/10.1145/2959100.2959103

Srivastava, A., Bala, P. K., & Kumar, B. (2020). New perspectives on gray sheep behavior in E-commerce recommendations. *Journal of Retailing and Consumer Services*, *53*(September 2018), 101764. https://doi.org/10.1016/j.jretconser.2019.02.018

Su, X., & Khoshgoftaar, T. M. (2009). A Survey of Collaborative Filtering Techniques. *Advances in Artificial Intelligence*, *2009*(Section 3), 1–19. https://doi.org/10.1155/2009/421425

Takács, G., & Tikk, D. (2012). Alternating least squares for personalized ranking. *RecSys'12 - Proceedings of the 6th ACM Conference on Recommender Systems*, 83–90. https://doi.org/10.1145/2365952.2365972

Tsai, C. F., & Hung, C. (2012). Cluster ensembles in collaborative filtering recommendation. *Applied Soft Computing Journal*, *12*(4), 1417–1425. https://doi.org/10.1016/j.asoc.2011.11.016

Tuch, A. N., Roth, S. P., Hornbæk, K., Opwis, K., & Bargas-Avila, J. A. (2012). Is beautiful really usable? Toward understanding the relation between usability, aesthetics, and affect in HCI. *Computers in Human Behavior*, *28*(5), 1596–1607. https://doi.org/10.1016/j.chb.2012.03.024

Van Rijsbergen, C. J. . (1979). Evaluation. *Information Retrieval*, 112–140. https://doi.org/10.1007/SpringerReference_16360

Volkovs, M., Yu, G., & Poutanen, T. (2017). DropoutNet: Addressing cold start in recommender systems. *Advances in Neural Information Processing Systems*, *2017-Decem*(Nips), 4958–4967.

Vozalis, M. G., & Margaritis, K. G. (2007). Using SVD and demographic data for the enhancement of generalized Collaborative Filtering. *Information Sciences*, *177*(15), 3017–3037. https://doi.org/10.1016/j.ins.2007.02.036

Walunj, S., & Sadafale, K. (2013). An online recommendation system for e-commerce based on Apache Mahout framework. *SIGMIS-CPR 2013 - Proceedings of the 2013 ACM Conference on Computers and People Research*, 153–158. https://doi.org/10.1145/2487294.2487328

Wei, P., Li, Y., Zhang, Z., Hu, T., Li, Z., & Liu, D. (2019). An optimization method for intrusion detection classification model based on deep belief network. *IEEE Access*, *7*, 87593–87605. https://doi.org/10.1109/ACCESS.2019.2925828

Wei, S., Ye, N., Zhang, S., Huang, X., & Zhu, J. (2012). Collaborative filtering recommendation algorithm based on item clustering and global similarity. *Proceedings of the 2012 5th International Conference on Business Intelligence and Financial Engineering, BIFE 2012*, *1*, 69–72. https://doi.org/10.1109/BIFE.2012.23

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1999*, 42–49. https://doi.org/10.1145/312624.312647

Yiming, Y., & Jan O., P. (1997). A Comparative Study on Feature Selection in Text Categorization. *Proceeding ICML '97 Proceedings of the Fourteenth International Conference on Machine Learning*, *53*(9), 412–420.

Yu, K., Schwaighofer, A., Tresp, V., Xu, X., & Kriegel, H. (2004). *Probabilistic Memory-Based Collaborative Filtering*. *16*(1), 56–69.

Yu, M., Quan, T., Peng, Q., Yu, X., & Liu, L. (2022). A model-based collaborate filtering algorithm based on stacked AutoEncoder. *Neural Computing and Applications*, *34*(4), 2503–2511. https://doi.org/10.1007/s00521-021-05933-8

Zhang, M., Tang, J., Zhang, X., & Xue, X. (2014). *Addressing cold start in recommender systems*. 73–82. https://doi.org/10.1145/2600428.2609599

Zhang, S., & Wu, X. (2011). Fundamentals of association rules in data mining and knowledge discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(2), 97–116. https://doi.org/10.1002/widm.10

Zheng, Y., Agnani, M., & Singh, M. (2017). Identifying grey sheep users by the distribution of user similarities in collaborative filtering. *RIIT 2017 - Proceedings of the 6th Annual Conference on Research in Information Technology*, 1–6. https://doi.org/10.1145/3125649.3125651

Zhu, Y., Shen, X., & Ye, C. (2016). Personalized Prediction and Sparsity Pursuit in Latent Factor Models. *Journal of the American Statistical Association*, *111*(513), 241–252. https://doi.org/10.1080/01621459.2014.999158