REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
MINISTERE DE L'ENSEIGNEMENT SUPERIEURE ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE IBN KHALDOUN - TIARET

MEMOIRE

Présenté à :

FACULTÉ MATHEMATIQUES ET INFORMATIQUE
DÉPARTEMENT D'INFORMATIQUE

Pour l'obtention du diplôme de :

MASTER

Spécialité : Réseaux et Télécommunication

Par :

- BAHLOUL Amine
- BELDJOUHER Ali

Sur le thème

Système de recommandation et réseau social : Exploiter les données d'annotation (les tags) pour les systèmes de recommandation

Soutenu publiquement à Tiaret devant le jury composé de :

Mr. MOSTEFAOUI Kadda	Grade Université MCA	Président
Mr. KOUADRIA Abderrahmane	Grade Université MAA	Encadreur
Mr. MEZZOUG Karim	Grade Université MAA	Examinateur

Remerciement

Nous remercions en premier lieu ALLAH qui nous a éclairé le chemin du savoir et qui nous a donné la volonté et la patience d'achever ce modeste travail de mémoire, notre grand salut sur le premier éducateur notre prophète Mohamed(satisfaction et salut de dieu soit sur lui).

Nous tenons à adresser nos remerciements à nos parents qui ont consenti des sacrifices et prodigué des encouragements tout au long de nos études.

Nous adressons nos vifs remerciements et nos sincères gratitudes à notre

Encadreur KOUADRIA Abderrahmane qui nous a fait l'honneur d'avoir la charge d'encadrer notre travail de mémoire avec grande patience pour la confiance qu'il a eue en notre projet et surtout pour ses orientations ainsi que son aide précieuse et ses conseils pour réaliser cette mémoire.

Nous remercions également notre jury d'avoir accepté de juger notre travail.

En fin, nous tenons à remercier également nos collègues pour leur aide à la réalisation de cette Modeste Mémoire.

Nous tenons à remercier nos professeurs de département d'informatique de nous avoir incités à travailler en mettant à notre disposition leurs expériences et leur compétence. On n'oublie surtout pas nos parents et nos proches pour leurs contributions, leurs soutiens et leur patience.

Merci à toutes et à tous :

Dédicaces

On dédie ce modeste travail :

À nos très chers parents.

À nos frères et sœurs.

Tous nos amis.

Et on remercie tous ceux et celles qui ont contribué à

réaliser ce travail.

Résumé.

Dans les systèmes de marquage social (tagging social), les utilisateurs ne fournissent pas de notation explicite sur les ressources (items) qui les intéressent. Au lieu d'évaluer les items par des notes, les utilisateurs annotent les ressources à l'aide de mots clés appelés tags, qui indiquent l'intérêt et la préférence de l'utilisateur. L'approche qu'on doit utiliser analysera le comportement de marquage de l'utilisateur et tentera d'estimer l'intérêt et la préférence de l'utilisateur. Le but de ce travail est donc l'utilisation d'un système de recommandation basé sur les données de marquage c'est-à-dire l'exploitation des informations d'annotations (tags) collaboratives fournies par les utilisateurs.

Mots-clés : Système de recommandation, profil utilisateur, Folksonomie, tag.

Table des matières

Résumé ..... **III**
Liste des figures ..... **VII**
Liste des tableaux ..... **IX**
Liste des abréviations ..... **X**
Introduction Générale **1**
Chapitre 1 : Systèmes de filtrage d'information et systèmes de recommandation **4**
1.1. Introduction 4
1.2. Les Systèmes de recommandation (SR) 4
1.2.1. Définition et fonctionnement 4
1.2.2. Les étapes principales de la recommandation 5
1.2.2.1. La collecte d'information 6
1.2.2.2. Modèle utilisateur 6
1.2.2.3. Liste de recommandations 7
1.3. Les techniques de recommandation 7
1.3.1. Filtrage basé sur le contenu 7
1.3.1.1. Principe 7
1.3.1.2. Avantages 8
1.3.1.3. Inconvénients 9
1.3.2. Filtrage collaboratif 10
1.3.2.1. Principe 10
1.3.2.2. Concepts de base 12
1.3.2.3. Avantages 14
1.3.2.4. Inconvénients 14
1.3.2.5. Techniques de filtrage collaboratif 15
1.3.2.5.1. Filtrage collaboratif basé sur la mémoire 15
1.3.2.5.2. Filtrage collaboratif basé sur un modèle 17
1.3.3. Filtrage hybride 21
1.4. Recommandation de ressources 23
1.5. Recommandation de tags 24
1.6. Conclusion 25
Chapitre 2 : Systèmes d'annotations (tagging) collaboratives **27**
2.1. Introduction 27
2.2. Systèmes de tagging 27

2.2.1. Définitions 27
2.2.2. Caractéristiques des systèmes de tagging 29
2.2.3. Propriétés d'un système du Tagging 29
2.2.4. Conception des systèmes de tagging 30
2.2.5. Motivations des utilisateurs 31
2.2.6. Structure d'une action de tagging 31
2.2.6.1. Structure en tripartie 31
2.2.6.2. Structure en tripartie avec lien inter ressources et inter utilisateurs 32
2.2.6.3. Structure en quadripartie 33
2.2.7. Quelques exemples de systèmes de tagging 34
2.2.8. Les limites des systèmes de tagging 34
2.3. Folksonomie 35
2.3.1. Définition 35
2.3.2. Caractéristiques 36
2.3.3. Type de folksonomies 37
2.3.4. Folksonomie Vs classification traditionnelle 37
2.3.5. Quelques règles pour une bonne indexation 38
2.3.6. Avantages, limites et futures perspectives 39
2.4. Travaux sur le tagging et folksonomie 40
2.4.1. Etudes sur la dynamique des systèmes de tagging 40
2.4.2. Etudes pour rapprochement des folksonomies et ontologies 42
2.4.3. Etudes pour exploitation des tags dans les systèmes de recommandation 43
2.5. Conclusion 43
Chapitre 3 : Présentation des systèmes de tagging existants. 46
3.1. Introduction 46
3.2. Systèmes de recommandation à base de tags 46
3.2.1. Détection des intérêts de l'utilisateur social 47
3.2.1.1. Le rôle du comportement social pour la détection des intérêts 48
3.2.1.2. Approches de détection des intérêts à partir du comportement social 49
3.2.1.2.1. Travaux basés sur les utilisateurs 49
3.2.1.2.1. Travaux basés sur les tags 52
3.2.1.2.2. Travaux basés sur les ressources 52
3.2.1.3. Approches d'enrichissement du profil utilisateur à partir du comportement social 53
3.2.1.3.1. Travaux basés sur les tags 53
3.2.1.3.1. Travaux basés sur les ressources 55

3.3. Conclusion....56
Chapitre 4 : Proposition et évaluation....58
4.1. Introduction....58
4.2. Notre système de recommandation basé sur les tags....58
4.3. Mesurer la similarité....59
4.4. Solution proposée :....61
4.4.1. Algorithme basé sur l'utilisateur....61
4.4.1.1. Similarité basée sur les tags (Tag-based similarity)....62
4.4.1.2. Similarité basée sur les items (Item-based similarity)....63
4.4.1.3. Similarité basée sur l'amitié (Friendship-based similarity)....64
4.4.1.4. Similarité globale....66
4.4.1.5. Génération de recommandations....66
4.4.2. Algorithme basée sur l'item....67
4.4.2.1. Similarité basée sur utilisateur-tag relation (user-tag Relationship similarity)68
4.4.2.2. Similarité basée sur les utilisateurs (User-based similarity)....68
4.4.2.3. Similarité basée sur les tags (tag based Similarity)....69
4.4.2.4. Similarité globale....70
4.4.2.5. Génération de recommandations....70
4.5. Dataset....71
4.6. Métrique d'évaluation....76
4.7. Mesures d'évaluation utilisée....77
4.8. Outils de développements utilisés....78
4.8.1. Python....78
4.8.2. Caractéristiques du langage....79
4.9. Présentation de l'application....79
4.10. EXPÉRIENCES ET ANALYSE....82
4.10.1. Trouver α et β....83
4.10.2. Résultats et analyse....86
4.11. Conclusion....93
Conclusion Générale....95
Bibliographie Et Références....98

Liste des figures

Figure 1. 1 - Facteurs de recommandation. 5
Figure 1. 2 - Filtrage basé sur le contenu 8
Figure 1. 3 - Filtrage collaboratif 11
Figure 1. 4 - Composantes d'un système de filtrage collaboratif 11
Figure 1. 5 - Processus de filtrage collaboratif 12
Figure 1. 6 - Valeurs possibles de la corrélation Pearson. 17
Figure 1. 7 - Algorithme des K-moyennes 19
Figure 1. 8 - Réseaux de Bayes 20
Figure 1. 9 - Le filtrage hybride. 21
Figure 2. 1 - Action de tagging en tripartie 32
Figure 2. 2 - Action de tagging en tripartie avec liens inter-ressources et inter-utilisateurs 33
Figure 2. 3 - Action de tagging en quadripartie 34
Figure 3. 1 - La dérivation du profil utilisateur selon les personnes 51
Figure 3. 2 - Enrichissement du profil utilisateur selon les tags des personnes proches 54
Figure 4. 1 - Similarité basée sur les tags 63
Figure 4. 2 - Similarité basée sur les items 64
Figure 4. 3- Pseudo-code permettant de calculer la similarité d'utilisateur sur la base d'informations d'amitié 65
Figure 4. 4 - Similarité basée sur l'amitié. 65
Figure 4. 5 - Approche basée sur l'utilisateur 67
Figure 4. 6 - Similarité basée utilisateur-tag relation 68
Figure 4. 7 - Similarité basée les utilisateurs 69
Figure 4. 8 - Similarité basée sur les tags 70
Figure 4. 9 - Approche basée sur l'item 71
Figure 4. 10 - Distribution du nombre de tag attribué par utilisateur dans la base de données Lastfm. 72
Figure 4. 11 - Distribution de tags pour les pistes 75
Figure 4. 12 - Division d'un ensemble de données en un ensemble d'apprentissage et un ensemble d'évaluation. 76
Figure 4. 13 - Code source read data in pandas. 80
Figure 4. 14 - Fenêtre l'ensemble de données. 80
Figure 4. 15 - Code source Division data in pandas. 80
Figure 4. 16 - Fenêtre Training Data. 81
Figure 4. 17 - Fenêtre Testing Training Data. 81

Figure 4. 18 - Code source convertie data en une matrice. 82
Figure 4. 19 - Fenêtre Matrice. 82
Figure 4. 20 - La précision sur l'algorithme basé sur l'utilisateur 87
Figure 4. 21 - Le Rappel sur l'algorithme basé sur l'utilisateur 87
Figure 4. 22- La précision sur l'algorithme basé sur l'item 89
Figure 4. 23 -Le Rappel sur l'algorithme basée sur l'item 89
Figure 4. 24 - Comparaison de la précision pour l'algorithme basé sur l'utilisateur 90
Figure 4. 25 - Comparaison de la précision pour l'algorithme basé sur l'item 91
Figure 4. 26 - Comparaison du rappel pour le modèle basée sur l'utilisateur 92
Figure 4. 27 - Comparaison du rappel pour l'algorithme basé sur l'item 92

Liste des tableaux

Tableau 1. 1 - Exemple de matrice d'usage..... 7
Tableau 1. 2 - Profil utilisateur sous forme d'un vecteur. 13
Tableau 1. 3 - Profil utilisateur sous forme matricielle.....13
Tableau 1. 4 - Matrice des évaluations attribuées aux documents par les utilisateurs. 13
Tableau 1. 5 - Matrice des votes { 1 , , … } : liste des ressources { 1 , , … }. 15
Tableau 1. 6 - Matrice des évaluations attribuées aux documents par les utilisateurs (rappel du tableau 1.4) 18
Tableau 1. 7 - Partition de la matrice des évaluations des utilisateurs. 19
Tableau 2.1 - Tableau comparatif des folksonomies et systèmes traditionnels d'indexation.38
Tableau 2. 2 - Attributs de conception des systèmes de tagging 41
Tableau 4. 1 - Statistiques de données de hetrec-lastfm-2k dataset.……73
Tableau 4. 2 - Exemple de fichier de tag 73
Tableau 4. 3 - Exemple de fichier d'utilisateur-artiste 74
Tableau 4. 4 - Exemple de fichier d'annotation (Utilisateurs-tags-artistes).....74
Tableau 4. 5 - Exemple de fichier d'utilisateur-utilisateur (ami).....74
Tableau 4. 6 - Valeur de précision pour les 5 top items. 84
Tableau 4. 7 - Valeur de rappel pour les 5 top items. 84
Tableau 4. 8 - Valeur de précision pour les 5 top items 85
Tableau 4. 9 - Valeur de rappel dans le top 5 85
Tableau 4. 10 -La précision sur l'algorithme basé sur l'utilisateur 86
Tableau 4. 11 - Le Rappel sur l'algorithme basé sur l'utilisateur 86
Tableau 4. 12 - La précision sur l'algorithme basé sur l'item.....88
Tableau 4. 13 - Le Rappel sur l'algorithme basé sur l'item.....88
Tableau 4. 14 - Comparaison de la précision 90
Tableau 4. 15 - Comparaison du rappel 91

Liste des abréviations

CF Filtrage collaboratif

TF-IDF Term Frequency-Inverse Document Frequency

IEML Information Economy Meta Language

CB Filtrage basé sur le contenu

DBLP Digital Bibliography & Library Project

Introduction Générale

Introduction Générale

Il existe une demande croissante en ce qui concerne les systèmes de recommandation et ce en raison de la surcharge d'informations sur le web. L'objectif de ces systèmes est de fournir des recommandations personnalisées de produits ou de services aux utilisateurs. Avec l'avènement du web social, le contenu généré par les utilisateurs a enrichi la dimension sociale du web. Les données de contenu fournies par l'utilisateur nous renseignent également sur l'utilisateur et il est possible d'apprendre ses préférences individuelles sur le web social. Cela ouvre des possibilités et des défis totalement nouveaux pour la recherche sur les systèmes de recommandation.

Les tags fournis par les utilisateurs constituent aujourd'hui un moyen populaire permettant aux utilisateurs d'organiser et de récupérer des items d'intérêt sur le web social. Le marquage social (Tagging Social) joue un rôle de plus en plus important à la fois sur les plates-formes web sociales tels que *last.fm*[1] et *YouTube*[2], ainsi que sur les sites de commerce électronique à grande échelle tels qu'Amazon[3]. Les applications web sociales encouragent les utilisateurs à partager et à classer en collaboration le contenu à l'aide de tags.

Actuellement, des travaux ont été réalisés sur la manière d'utiliser les informations de marquage collaboratif pour recommander des balises personnalisées aux utilisateurs **[78]**, cependant peu de travaux ont été réalisés sur l'utilisation des informations de marquage pour aider les utilisateurs à trouver facilement et rapidement les items intéressés. En particulier, les techniques de filtrage collaboratif couramment utilisées actuellement ne fonctionnent pas bien avec la relation tridimensionnelle distincte entre les utilisateurs, les tags et les items. Ainsi, comment recommander aux utilisateurs des items personnalisés en fonction des informations de marquage devient une question de recherche importante et la recherche ne fait que commencer.

Ce travail vise à traiter la manière dont les données d'annotation (les tags) fournies par l'utilisateur peuvent être utilisées pour créer de meilleurs systèmes de recommandation. Un algorithme de recommandation basé sur les tags sera proposé, il exploite les données de marquage fournies par l'utilisateur et produisent des recommandations plus précises. Sur la base de cette idée, on doit montrer comment les tags peuvent être utilisés pour expliquer à l'utilisateur les recommandations générées automatiquement sous une forme claire, compréhensible et de manière intuitive. Par conséquent, nous avons proposé l'utilisation d'une similarité globale qui non seulement prend en compte les activités de marquage des utilisateurs, mais intègre également leurs relations sociales, telles que les amitiés, afin de trouver les plus proches des

[1] www.last.fm
[2] www.youtube.com
[3] www.amazon.com

voisins de l'utilisateur active. Les résultats expérimentaux obtenus sur l'ensemble de données Last.fm montrent des résultats positifs en ce qui concerne notre approche proposée.

Enfin notre mémoire est structuré en quatre chapitres :

- **Chapitre I :** Nous présentons au sein de ce chapitre un état de l'art sur les systèmes de recommandation, tels que leurs principes, fonctionnements et caractéristiques, ainsi que nous énonçons les différents types de filtrage.
- **Chapitre II :** Dans ce chapitre nous allons définir les systèmes de tagging, les folksonomies qui en découlent et nous détaillons les concepts de base liés à ces paradigmes.
- **Chapitre III :** Le troisième chapitre concerne les systèmes de recommandation à base de tags ainsi que les travaux de détection des intérêts de l'utilisateur social. Ces travaux analysent le comportement social de l'utilisateur et plus précisément son comportement d'annotation.
- **Chapitre IV** : Dans ce quatrième et dernier chapitre, nous mettons en relief notre proposition qui essaie de répondre à la problématique posée.

Chapitre I

Systèmes de filtrage d'information et systèmes de recommandation

Chapitre 1 : Systèmes de filtrage d'information et systèmes de recommandation

1.1. Introduction

Avec la très grande masse d'information, aujourd'hui disponible sur l'Internet et les besoins en communication, en échange d'idées et en partage d'informations ce qui a nécessité l'apparition de nouvelles fonctions et applications sur les réseaux sociaux, il est devenu primordial de concevoir des mécanismes qui permettent aux utilisateurs d'accéder aux choses qui les intéresse le plus rapidement possible.

Delà, le système de recommandation est né dont sa qualité est étroitement liée à sa capacité à prendre en compte et à traiter une grande quantité d'évaluation.

Les systèmes de recommandation sont nés de la volonté de pallier au problème de surcharge d'information sur le Web, combinant ainsi des techniques de filtrage d'information, personnalisation, intelligence artificielle, réseaux sociaux et interaction personne-machine, les systèmes de recommandation fournissent à des utilisateurs des suggestions qui répondent à leurs besoins et préférences informationnelles. En effet, les systèmes de recommandation sont particulièrement sollicités dans les applications de commerce électronique.

1.2. Les Systèmes de recommandation (SR)

Dans cette section, nous entreprenons en premier lieu par définir les Systèmes de recommandation. En deuxième lieu, nous expliquons les différentes techniques de recommandation. En troisième lieu, nous détaillons les travaux de recommandations selon le type de données à recommander, suivant l'objectif visé et bien évidemment selon les informations disponibles. Cette dernière peut être un tag, une ressource ou une personne.

1.2.1. Définition et fonctionnement

En raison de la popularité croissante des systèmes de recommandation dans de nombreux domaines, ce qui a conduit à l'apparition d'un conflit dans sa définition. Il existe de nombreuses définitions, parmi elles la définition de *Saimadhu* **[1]**, qui propose la définition suivante :

" *Un système de recommandation est une boîte noire qui analyse un certain ensemble d'utilisateurs et présente les items où un seul utilisateur peut aimer* ". La définition la plus générale de *Robin Burke* **[2]** qui les défini comme étant : "*Des systèmes capable de fournir des recommandations personnalisées permettant de guider l'utilisateur vers des ressources intéressantes et utiles au sein d'un espace de données important* ".

On relève au sein de ces deux définitions que les systèmes de recommandation fonctionnent idéalement sur les manières suivantes :

Soit sur les propriétés des éléments que l'utilisateur aime ou non " les items", soit sur la similarité entre les utilisateurs et de leur recommander des items en conséquence. Il est également possible de combiner ces deux méthodes pour construire un moteur beaucoup plus robuste de recommandation. Beaucoup d'algorithmes ont été utilisés pour mesurer la similarité des utilisateurs ou des items dans les systèmes de recommandation on cite par exemple le k-nearest neighbors (k-NN) et Pearson corrélation ...

Un système de recommandation prend en compte plusieurs facteurs en considération pour faire une recommandation à un utilisateur soit :

- ***Le profil de l'utilisateur :*** Age, situation géographique, historique, ...
- ***Informations sur les déférents items disponibles :*** Contenu associé à l'item.
- ***Les interactions des utilisateurs :*** contenu de navigation
- ***Le contexte dans lequel les items seront affichés :*** Sous-catégorie d'items qui doivent être considéré

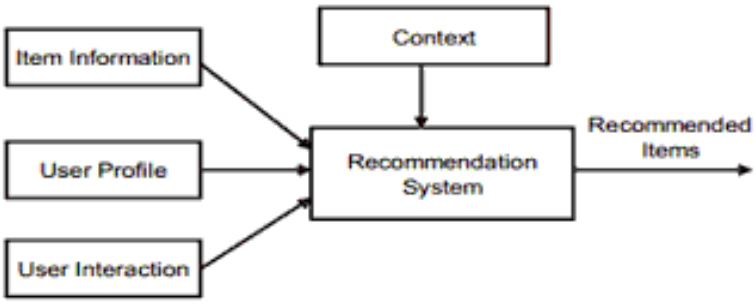

Figure 1. 1 - Facteurs de recommandation.

1.2.2. Les étapes principales de la recommandation

Un système de recommandation requiert généralement trois (03) étapes :

1. La première consiste à recueillir de l'information sur l'utilisateur ;
2. La deuxième consiste à bâtir une matrice ou un modèle utilisateur contenant l'information recueillie ;
3. La troisième consiste à extraire à partir de cette matrice une liste de recommandations.

1.2.2.1. La collecte d'information

Pour être pertinent, un système de recommandation doit pouvoir faire des prédictions sur les intérêts des utilisateurs. Il faut donc pouvoir collecter un certain nombre de données sur ceux-ci, afin d'être capable de construire un profil pour chaque utilisateur. Une distinction peut être faite entre deux (02) formes de collecte de données :

a) Collecte de données explicites - Filtrage actif : La collecte repose sur le fait que l'utilisateur indique explicitement au système ses intérêts.

Exemple : Demander à un utilisateur de commenter, taguer/étiqueter, noter, liker ou encore ajouter comme favoris des contenus (objets, items...) qui l'intéressent. On utilise souvent une échelle de ratings allant de 1 étoile (je n'aime pas du tout) à 5 étoiles (j'aime beaucoup) qui sont ensuite transformées en valeurs numériques, afin de pouvoir être utilisées par les algorithmes de recommandation.

Avantage : Capacité à reconstruire l'historique d'un individu et capacité à éviter d'agréger une information qui ne correspond pas à cet unique utilisateur (plusieurs personnes sur un même poste).

Inconvénient : Les informations recueillies peuvent contenir un biais dit de déclaration **[3]**.

b) Collecte de données implicite - Filtrage passif : Elle repose sur une observation et une analyse des comportements de l'utilisateur effectués de façon implicite dans l'application qui embarque le système de recommandation, où le tout se fait en "arrière-plan" (en gros sans rien demander à l'utilisateur).

Avantage : Aucune information n'est demandée aux utilisateurs, du fait que toutes les informations sont collectées automatiquement. Les données récupérées sont à priori justes et ne contiennent pas de biais de déclaration.

Inconvénient : Les données récupérées sont plus difficilement attribuables à un utilisateur et peuvent donc contenir des biais d'attribution (utilisation commune d'un même compte par plusieurs utilisateurs). Un utilisateur peut ne pas aimer certains livres qu'il les a achetés, ou il peut les avoir acheté pour quelqu'un d'autre.

1.2.2.2. Modèle utilisateur

Le modèle utilisateur se présente généralement sous forme de matrice appelée " matrice d'usages". On peut les représenter sous forme d'un tableau qui contient des données recueillies sur l'utilisateur associées aux produits disponibles sur le site web.

	Item 1	Item 2	Item 3	Item 4	Item 5	...
User 1	😎			😎		...
User 2		😎	😎			...
User 3		😎				...
User 4			😎	😎	😎	...
...	...	...	...	...	...	...

Tableau 1. 1 - Exemple de matrice d'usage

Le tableau 1.1 présente un exemple fictif de matrice binaire contenant des informations de type " l'utilisateur u a apprécié n'a pas apprécié l'item i ". Ces informations peuvent également être " à acheter/n'à pas acheter ", "à consulté/n'à pas consulter ", etc. Elles peuvent également se mesurer sur un nombre plus élever de classes : " a mis 1/2/3/4/5 étoiles " etc. Un autre point important c'est comment le temps influence sur le profil de l'utilisateur. Les intérêts des utilisateurs évoluent en général en cours du temps. Les données du modèle utilisateurs devraient donc constamment être réajustées pour rester conformes aux nouveaux centres d'intérêts de l'utilisateur.

1.2.2.3. Liste de recommandations

Pour extraire une liste de suggestions à partir d'un modèle utilisateur, les algorithmes utilisent la notion de mesure de similarité entre objets ou personnes décrit par le modèle utilisateur. La similarité a pour but de donner une valeur ou un nombre (au sens mathématique du terme) à la ressemblance entre 2 choses. Plus la ressemblance est forte, plus la valeur de la similarité sera grande. A l'inverse, plus la ressemblance est faible, et plus la valeur de la similarité sera petite. On verra plus tard quelques exemples au sein de ce texte.

1.3. Les techniques de recommandation

Il est possible de classer les systèmes de recommandation selon trois principales approches, basées sur le filtrage collaboratif, approches basées sur le contenu et les approches hybrides. Nous détaillons ces approches dans ce qui suit.

1.3.1. Filtrage basé sur le contenu

1.3.1.1. Principe

Le filtrage basé sur le contenu (Content-based Filtering), qui est une évolution générale des études sur le filtrage d'information, s'appuie sur le contenu des documents (thèmes abordés) pour les comparer à un profil lui-même constitué de thèmes. Chaque utilisateur du système

possède alors un profil qui décrit ses propres centres d'intérêt **[4]**. Deux fonctionnalités centrales ressortent, pour un système de filtrage :

- La sélection des documents pertinents vis-à-vis du profil.
- La mise à jour du profil en fonction du retour de pertinence fourni par l'utilisateur sur les documents qu'il a reçus. La mise à jour se fait par intégration des thèmes abordés dans les documents jugés pertinents **[5]**.

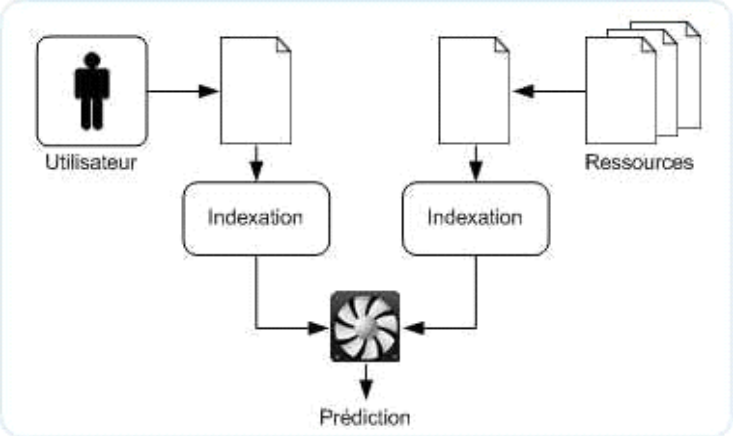

Figure 1. 2 - Filtrage basé sur le contenu

Le déroulement du processus de recommandation se résume de la manière suivante :

1. Extraction d'informations : Dans un premier temps, le système interroge (*Amazon* par exemple), sur une thématique particulière et obtient une liste de livres. L'information pertinente du livre est extraite à partir du titre de l'ouvrage et le profil de l'auteur en utilisant des patrons.

2. Evaluation utilisateur : L'utilisateur doit donner une note de 1 à 10 aux livres proposés par le système.

3. Apprentissage de profils : Les choix effectués par les différents utilisateurs sont appris par un classificateur Bayesien.

4. Recommandation : Le système propose de nouveaux livres ordonnés en termes d'adéquation avec les profils des utilisateurs.

1.3.1.2. Avantages

Les systèmes de recommandation basés sur le contenu présentent les avantages suivants :

- ✓ Le premier avantage du filtrage basé sur le contenu c'est qu'il peut répondre aux intérêts à long terme des utilisateurs, en employant des techniques efficaces dans le domaine de l'intelligence artificielle pour la mise à jour des profils ainsi que le

recoupement entre profils et documents. En outre, l'utilisateur dans un tel système ne dépend absolument pas des autres **[4]**.

- ✓ Recommander des items similaires à ceux que les utilisateurs ont aimés dans le passé. Le profil des utilisateurs est la clé.
- ✓ Le matching entre les préférences de l'utilisateur et les caractéristiques des items fonctionne aussi pour les données textuelles.
- ✓ Nous n'avons pas besoin de données sur les autres utilisateurs.
- ✓ Absence de problème de faible densité.
- ✓ Possibilité de faire recommander de nouveaux items ou même des items qui ne sont pas populaires **[6]**.

1.3.1.3. Inconvénients

- ✗ La capacité à traiter d'autres critères de pertinence que les critères strictement thématiques posent également problème. Le filtrage des documents basé sur le contenu ne permet pas d'intégrer d'autres facteurs de pertinence que le facteur thématique.

  Pourtant il existe de nombreux autres facteurs de pertinence comme par exemple l'adéquation entre le public visé par l'auteur et l'utilisateur, ou encore la qualité scientifique des faits présentés, la fiabilité de la source d'information, le degré de précision des faits présentés, etc.
- ✗ L'effet entonnoir : Le profil utilisateur évolue naturellement par restriction progressive sur les thèmes recherchés. Ainsi, l'utilisateur ne reçoit que les recommandations relatives aux thèmes présentés dans son profil, une fois devenu stable. Par conséquent, il ne peut pas découvrir de nouveaux domaines potentiellement intéressants pour lui. Par exemple, lorsqu'un nouvel axe de recherche surgit dans un domaine, avec de nouveaux termes pour décrire les nouveaux concepts, ces termes n'apparaissent pas dans le profil, ce qui élimine automatiquement les documents par filtrage, l'utilisateur n'aura donc jamais l'occasion d'exprimer un retour de pertinence positif envers ce nouvel axe de recherche, à moins d'en avoir connaissance par ailleurs et de modifier son profil manuellement en ajoutant les termes pertinents **[5]**.
- ✗ Tous les contenus ne peuvent pas être représentés avec des mots-clés (exemple : les images,…).

- Des items représentés par le même ensemble de mots-clés ne peuvent pas être distingués.
- Les utilisateurs avec des milliers d'achats/items sont un problème.
- Nouvel utilisateur : Pas d'historique.
- "Over- specialization " : Limitation aux items similaires.
- Pour produire des recommandations précises, l'utilisateur doit fournir un " feedback " sur les suggestions - les utilisateurs.
- Entièrement basé sur les scores d'items et de sujets d'intérêt : Moins il y a de scores, plus l'ensemble de recommandations possibles est limité.
- La difficulté d'indexation des documents multimédias: La croissance des documents multimédias (texte, image, vidéos, etc.) pose le problème de la prise en compte de l'information structurelle des documents pour aider à identifier les contenus multimédias pertinents.
- L'effet de masse: L'utilisateur ne bénéficie pas des jugements que d'autres utilisateurs peuvent faire sur les documents qu'il reçoit. L'utilisateur doit procéder lui-même à l'analyse des documents reçus, analyse qui fait intervenir d'autres critères que celui de la thématique **[7]**.
- Filtrage basé sur le critère thématique uniquement, absence d'autres facteurs comme la qualité scientifique, le public visé, l'intérêt porté par l'utilisateur, etc.
- Problème de démarrage à froid : Un nouvel utilisateur du système éprouve des difficultés à exprimer son profil en spécifiant des thèmes qui l'intéressent. Ceci malgré les techniques d'apprentissage ou l'utilisateur fournit des textes exemples **[8]**.

1.3.2. Filtrage collaboratif

1.3.2.1. Principe

Le filtrage collaboratif se base sur l'hypothèse que les gens à la recherche d'information devraient pouvoir se servir de ce que d'autres ont déjà trouvé et évalué. Cette approche résout les problèmes de l'approche basée sur le contenu sémantique ; il devient possible de traiter n'importe quelle forme de contenu et de diffuser des ressources non nécessairement similaires à celles déjà reçues. Pour ce faire, pour chaque utilisateur d'un système de filtrage collaboratif, un ensemble de proches voisins est identifié, et la décision de proposer ou non un document à un utilisateur dépendra des appréciations des membres de son voisinage.

Le filtrage collaboratif emploie des méthodes statistiques pour faire des prévisions basées sur des configurations des intérêts des utilisateurs. Ces prévisions sont exploitées pour faire des

propositions à un utilisateur individuel, en se fondant sur la corrélation entre son propre profil personnel et les profils d'autres utilisateurs qui présentent des intérêts et goûts semblables.

Le principe simplifié de ces approches est "les clients qui ont acheté un produit x ont aussi acheté le produit y". Ces approches ont connu un succès dans le domaine de e-commerce surtout avec *amazon*, dans les médias sociaux comme MovieLens et *LastFm* **[9]** ont introduit un algorithme de filtrage collaboratif basé sur les utilisateurs nommé "user-based CF" qui analyse des similarités entre les utilisateurs pour la création d'ensembles d'utilisateurs les plus proches possible d'un utilisateur donné. *Sarwar* **[10]** ont introduit un algorithme de filtrage collaboratif basé sur les items (produits) nommé "item-based CF" qui analyse des similarités entre les items pour la création d'ensembles d'utilisateurs les plus proches (les plus similaires) possible d'un utilisateur donné.

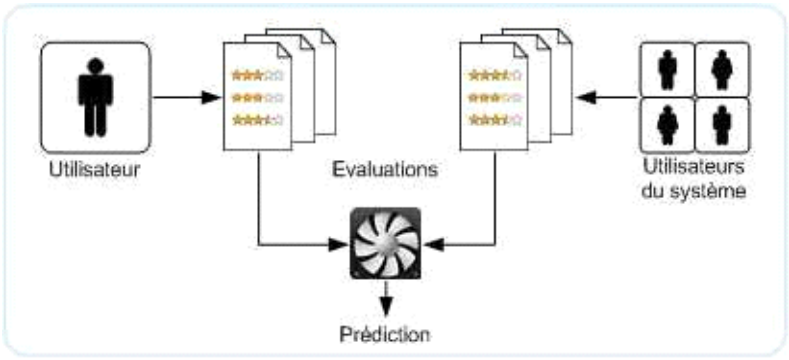

Figure 1. 3 - Filtrage collaboratif

Un système de filtrage collaboratif est organisé comme suit :

a) Collecter les appréciations et le comportement des utilisateurs. En général l'utilisateur fournit des évaluations sous forme de notes, sur un ou plusieurs axes : qualité, correspondance au besoin, etc.
b) Intégrer ces informations au profil de l'utilisateur.
c) Le système utilise ces informations pour faire des recommandations.

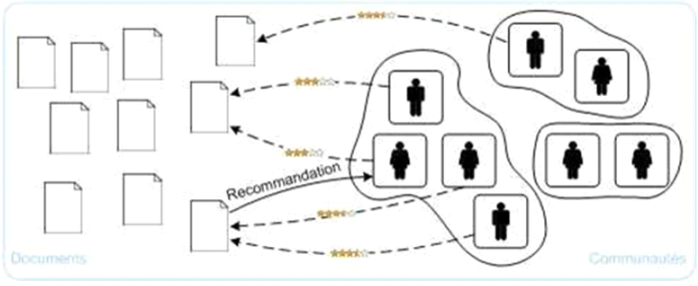

Figure 1. 4 - Composantes d'un système de filtrage collaboratif

L'utilisateur peut alors demander au système de :

- Lui suggérer une ressource susceptible de lui plaire.
- Le prévenir des ressources qu'il ne devrait pas apprécier.
- Donner une estimation d'évaluation d'une certaine ressource.

Le schéma ci-dessus *Berrut* **[11]** montre l'architecture globale d'un système de filtrage collaboratif.

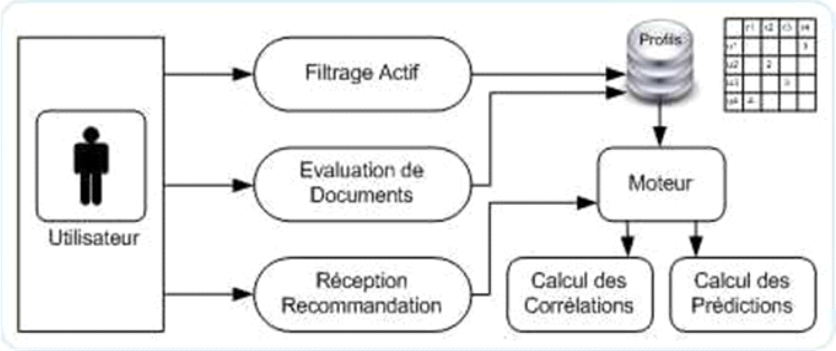

Figure 1. 5 - Processus de filtrage collaboratif

1.3.2.2. Concepts de base

Evaluation : Formalisation du goût d'un utilisateur pour un document. L'évaluation peut être « explicite » : Cela signifie qu'elle est donnée par l'utilisateur sous la forme d'une note sur une échelle donnée, de 1 à 5 par exemple. Elle peut aussi être «implicite», et dans ce cas c'est le système qui interprète certaines actions ou certains comportements de l'utilisateur comme des évaluations : Par exemple le fait d'imprimer un document, de le télécharger, ou de passer certain temps à le lire **[12]**.

Prédiction : Estimation que le système fait de l'intérêt qu'un document présente pour un utilisateur. C'est sur la base de la prédiction que le document est recommandé ou non ; la prédiction tient compte des évaluations des autres sur ce document, et de leur proximité avec l'utilisateur à qui est destinée la recommandation **[12]**.

Recommandation : Choix des documents dont la valeur de prédiction calculée est considérée pertinente pour l'utilisateur en question. Pour réaliser une telle sélection, un seuil de pertinence est défini **[12]**.

Profil : Le profil utilisateur est défini par un vecteur de dimension (correspondant au nombre de ressources disponibles). Chaque valeur du vecteur représente, soit l'évaluation que l'utilisateur a attribuée à la ressource, soit une valeur par défaut spécifiant que l'utilisateur n'a pas évalué la ressource.

Ressources ()	r_1	r_2	r_3	r_4	r_5
Utilisateur ()	4	-	2	-	7

Tableau 1. 2 - Profil utilisateur sous forme d'un vecteur.

Le profil utilisateur peut aussi se représenter sous la forme de matrice booléenne. Dans ce cas, nous ne considérons plus les évaluations mais les classes d'évaluation du type « Aime », « N'aime pas » correspondant à des intervalles de valeurs d'évaluations (par exemple, {0, 1, 2, 4} correspond à la classe « N'aime pas » et {4, 5, 6, 7} à la classe « Aime »). Le profil s'exprime alors sous la forme d'une matrice(c × m), où est le nombre de classes considérées (en général c = 2) et m le nombre de ressources disponibles. Les valeurs booléennes pour une matrice spécifient l'appartenance à une classe pour la ressource considérée **[13]**.

Ressources ()	r_1	r_2	r_3	r_4	r_5
« Aime »	1	0	0	0	1
«N'aime pas»	1	0	1	0	0

Tableau 1. 3 - Profil utilisateur sous forme matricielle.

Base de profil : Ensemble de profils stockés dans la base de données du système de filtrage collaboratif **[13]**.

La base de profil se présente sous la forme d'un tableau (matrice) dont chaque ligne correspond à un profil utilisateur (tableau 1.4).

	r_1	r_2	r_3	r_4	r_5
u_1	-	-	7	6	-
u_2	-	-	5	6	7
u_3	-	-	**6**	6	7
u_4	7	5	-	-	7
u_5	7	6	-	-	7

Tableau 1. 4 - Matrice des évaluations attribuées aux documents par les utilisateurs.

Cette matrice correspond à un ensemble d'évaluations $eval(u_i, r_j)$ de l'utilisateur (u_i) sur le document (r_j). Par exemple : $eval(u3, r3) = 6$ représente l'évaluation de l'utilisateur $(u3)$ sur le document $(r3)$.

1.3.2.3. Avantages

- ✓ Grace à son indépendance vis-à-vis de la représentation des documents, le filtrage collaboratif permet de résoudre les problèmes liés au filtrage basé sur le contenu, et donc de filtrer tout type d'information (textes, images, vidéos).
- ✓ L'effet entonnoir est éliminé par le filtrage collaboratif. Pour qu'un utilisateur reçoive un document, il suffit qu'un autre utilisateur de profil proche l'ait jugé intéressant, et cela quelques soient les termes qui indexent le contenu du document. L'utilisateur peut alors ouvrir son profil sur un nouveau thème en donnant simplement un retour de pertinence positif sur ce document.
- ✓ Un autre avantage du filtrage collaboratif est que les jugements de valeur des utilisateurs intègrent non seulement la dimension thématique mais aussi d'autre facteurs relatifs à la qualité de documents tels que la diversité, la nouveauté, l'adéquation du public visé, etc.

1.3.2.4. Inconvénients

De nombreux systèmes de recommandation s'appuient sur le filtrage collaboratif **[4]**, en raison des avantages importants ci-dessus. On constate néanmoins certains inconvénients de cette technique :

- × « Le démarrage à froid »: Dans un système de filtrage d'information collaboratif, le démarrage à froid apparaît à chaque nouvel utilisateur et nouveau document. En effet, le profil d'un nouvel utilisateur est inexistant et sa communauté est encore inconnue, ce qui conduit à l'impossibilité de fournir des recommandations pertinentes. Ce problème peut être réduit en demandant à l'utilisateur d'évaluer un ensemble de documents que le système lui présente. De la même façon un nouveau document ne peut être recommandé, car il n'est évalué par aucun utilisateur. Ce problème est généralement traité en combinant une approche de filtrage basé sur le contenu avec l'approche collaborative par exemple utilisant la similarité entre documents.
- × « La masse critique » : Afin de former de meilleures communautés, le système exige un nombre suffisant d'évaluations en commun entre les utilisateurs pour les comparer entre eux. Par exemple, on ne peut pas conclure que deux personnes sont dans une même communauté si elles n'ont qu'une seule évaluation en commun et par conséquent les personnes ayant des goûts peux fréquents risquent de ne pas recevoir de propositions.

- « Rapport coût-bénéfice » : Pour l'utilisateur d'un système de filtrage collaboratif, le rapport coût (son effort d'évaluation) et le bénéfice (les documents reçus automatiquement) varie au cours du temps. En particulier, au début de l'utilisation du système, ce rapport lui est souvent défavorable, ce qui peut le décourager d'utiliser le système pour atteindre une phase plus favorable. La défection des utilisateurs pénalise alors l'ensemble des performances du système, qui ne fonctionne bien qu'avec une participation active d'un nombre suffisant d'utilisateur.

1.3.2.5. Techniques de filtrage collaboratif

L'exploitation des données disponibles dans un système de filtrage peut se faire de plusieurs manières. *Breese et al*. proposent une classification intéressante des techniques de filtrage collaboratif : Les algorithmes basés « mémoire » et les algorithmes basés « modèle » **[5]**.

1.3.2.5.1. Filtrage collaboratif basé sur la mémoire

Le filtrage collaboratif basé sur la mémoire utilise une matrice des votes contenant les préférences des utilisateurs pour prédire des sujets additionnels ou des produits auxquels un nouvel utilisateur peut être s'intéressé. L'objectif d'un filtrage collaboratif basé sur la mémoire est de prédire l'utilité des ressources (items) pour un utilisateur particulier (l'utilisateur actif) basé sur la base des votes d'utilisateur.

	r_1	r_2	r_3	…	r_m
u_1	5	2	3	…	4
u_2	4	1	1	…	0
u_3	4	0	2	…	3
…	…	…	…	…	…
u_4	4	3	5	…	5

Tableau 1. 5 - Matrice des votes { 1 , , … } : liste des ressources { 1 , , … } : liste des utilisateurs.

La base de données des votes d'utilisateur contient un ensemble des votes , correspondant au vote de l'utilisateur i sur la ressource j. Si est l'ensemble des items évalués par l'utilisateur i, alors l'évaluation moyenne pour l'utilisateur i est définie comme :

$$\bar{v}_i = \frac{1}{|l_i|}\sum_{j \in l_i} v_{i,j} \quad (1.1)$$

L'évaluation prédit sur l'item j pour l'utilisateur actif a est une somme pondérée des évaluations des autres utilisateurs:

$$p_{a,j} = \bar{v}_a + k\sum_{i=1}^{n} w(a,i)\left(v_{i,j} - \bar{v}_a\right) \quad (1.2)$$

Avec :

$v_{i,j}$: L'évaluation de la ressource j par l'utilisateur i.

$\overline{v_a}$: La moyenne de l'ensemble des évaluations fournies par l'utilisateur i.

W (a, i) : Le coefficient de pondération liant l'utilisateur actif a et i. Il peut refléter la distance, corrélation ou similarité entre deux utilisateurs a et i.

K : Un coefficient de normalisation, il définit comme suit :

$$k = \frac{1}{\sum_i^n w(a,i)} \quad (1.3)$$

n : Le nombre des utilisateurs dont le poids n'est pas égal à zéro.

K est un coefficient de normalisation permettant d'harmoniser les votes afin de minimiser l'influence des utilisateurs ayant tendance à noter de façon extrême (uniquement des notes très élevées ou très basses). Par ailleurs, le coefficient de pondération, $w(a,i)$ représente la similarité existante entre l'utilisateur actif et les autres. Plus ils sont proches et plus le coefficient est grand. Ce coefficient peut être défini à partir de la distance, de la corrélation ou de la similarité entre chaque utilisateur i et l'utilisateur actif.

Corrélation de Pearson

Le coefficient de Pearson a été utilisé dans le contexte des systèmes de filtrage collaboratif pour la première fois dans les recherches du *GroupLens Research*[4] **[9]**.

$$w(a,i) = \frac{\sum_j \left(v_{a,j} - \bar{v}_a\right)\left(v_{i,j} - \bar{v}_i\right)}{\sqrt{\sum_j \left(v_{a,j} - \bar{v}_a\right)^2 \sum_j \left(v_{i,j} - \bar{v}_i\right)^2}} \quad (1.4)$$

[4] www.grouplens.org

Dans ce cas, la valeur, w (a, i) est entre -1 et 1. Si cette valeur est égale à -1, nous disons que les deux utilisateurs sont fortement corrélés opposés. Si la valeur est égale à 1, les deux utilisateurs sont fortement corrélés semblables. Les deux utilisateurs sont considérés être indépendants si la valeur est égale à 0.

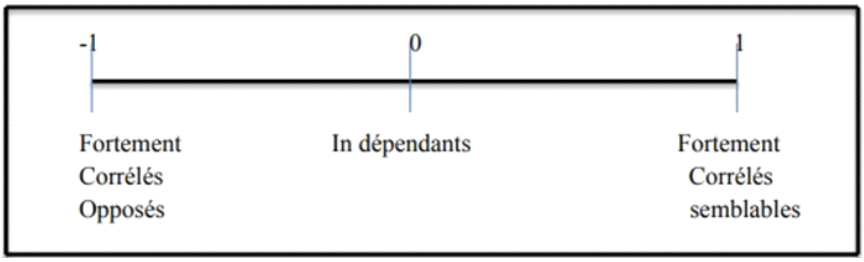

Figure 1. 6 - Valeurs possibles de la corrélation Pearson.

Cosinus des vecteurs

Pour le cas de la similarité vectorielle, ces poids sont définis selon la formule suivante:

$$w(a,i) = \sum_{j} \frac{v_{a,j}}{\sqrt{\sum_{k \in I_a} v_{a,k}{}^2}} \frac{v_{i,j}}{\sqrt{\sum_{k \in I_i} v_{i,k}{}^2}} \qquad (1.5)$$

De tels algorithmes ont l'avantage d'être simples à mettre en œuvre et d'évoluer dynamiquement en fonction des profils utilisateurs. En effet, toute évolution d'un utilisateur se répercute directement dans le calcul de prédiction.

Cependant, ces algorithmes souffrent de deux inconvénients majeurs. D'une part, la forte complexité combinatoire empêche le passage à l'échelle pour un nombre important d'utilisateurs et de ressources. D'autre part, le faible nombre de ressources communément évaluées par les utilisateurs engendre des prédictions peu pertinentes **[5]**.

1.3.2.5.2. Filtrage collaboratif basé sur un modèle

Le deuxième type d'algorithmes, est comme le nom l'indique basés sur des modèles, supposés réduire la complexité. Ces modèles peuvent être probabilistes et utiliser l'espérance de l'évaluation pour calculer la prédiction. Comme ils peuvent être basés sur des classificateurs permettant de créer des classes pour réduire la complexité.

Les algorithmes basés « modèle » utilisent la base de données des évaluations des utilisateurs pour estimer ou apprendre un modèle qui est alors utilisé pour les prédictions.

Du point de vue probabiliste, la tâche de prédiction d'une évaluation peut être vue comme le calcul de la valeur espérée d'une évaluation, étant donné ce que l'on sait d'un utilisateur.

Nous présentons ici deux modèles probabilistes : Le modèle à base de clusters et le modèle à base de réseau bayésien.

Modèle à base de clusters

Un cluster (groupe) est un ensemble d'objets qui sont « similaires » entre eux et « différents » des objets appartenant aux autres groupes **[14]**.

L'idée du modèle à base de cluster est de regrouper en clusters (en groupes) les utilisateurs ayant les mêmes goûts et de regrouper en clusters les documents portant sur les mêmes sujets ou qui ont tendance à plaire aux mêmes personnes. Ainsi pour prédire l'évaluation qu'un utilisateur donnera à un document, on pourra utiliser les évaluations des utilisateurs qui appartiennent à son groupe. En d'autres termes, on veut associer une classe à chacun des utilisateurs, ainsi qu'à chacun des documents. Ces classes étant à priori inconnues, elles doivent être dérivées du processus d'estimation du modèle.

Pour obtenir un bon partitionnement, il convient de maximiser la similarité des observations à l'intérieur d'un cluster et minimiser la similarité entre clusters.

En entrée, ce que nous avons, c'est l'ensemble d'enregistrement de : qui a apprécié quoi, représenté par une matrice d'évaluations des documents par les utilisateurs (tableau1.6). L'idée du clustering est alors de former des blocs les plus pertinents et significatifs possible à partir dès cette matrice (tableau1.6), et former ainsi des clusters d'utilisateurs et des clusters de documents, tels que les utilisateurs considérés comme similaires tendent à noter de la même façon les documents considérés comme similaires.

	r_1	r_2	r_3	r_4	r_5
u_1	-	-	7	6	
u_2	-	-	5	6	7
u_3	-	-	6	6	7
u_4	7	5	-	-	7
u_5	7	6	-	-	7

Tableau 1. 6 - Matrice des évaluations attribuées aux documents par les utilisateurs (rappel du tableau 1.4)

	r_1	r_2	r_3	r_4	r_5
u_1			7	6	
u_2			5	6	7
u_3			6	6	7
u_4	7				7
u_5	7				7

Tableau 1. 7 - Partition de la matrice des évaluations des utilisateurs.

Puis la partition est évaluée en estimant, pour chaque bloc formé, la probabilité que les utilisateurs d'un même cluster apprécient les documents d'un même cluster, le but étant que ces probabilités soient les plus proches possible de 1, c'est-à-dire que tout le monde soit du même avis dans le groupe.

Il existe plusieurs algorithmes qui permettent la classification des utilisateurs et des ressources, nous présentons dans ce qui suit l'algorithmes des *K-moyennes* et *RecTree* :

- **K- moyennes**

L'idée principale de l'algorithme des K-moyennes est de classifier des objets (utilisateurs, documents) en K classes en minimisant la variance intra-classe et en maximisant l'écartement interclasses. L'algorithme des K-moyennes se décompose en quatre (04) étapes :

1. Choisir aléatoirement K objets pour former les K clusters initiaux. Ces objets représentent entre autres les centres de clusters, ne contenant qu'un seul élément.
2. Réaffecter les objets à un cluster. Chaque objet est assigné à la classe dont il est plus proche du centre, selon une mesure de distance telle que la distance euclidienne.
3. Recalculer les nouveaux centres des K clusters.
4. Répéter les étapes 2 et 3 jusqu'à ce que plus aucune réaffectation ne soit possible.

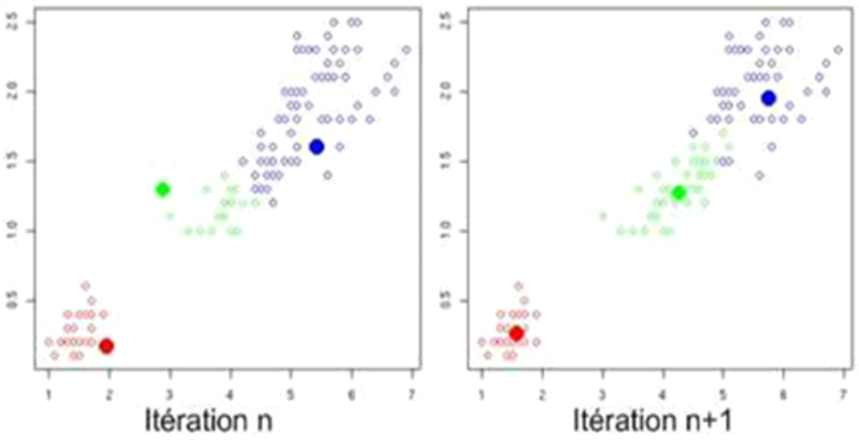

Figure 1. 7 - Algorithme des K-moyennes

Algorithme des K-moyennes

K_moyennes ($Object = \{X_1, X_2, \dots, X_n\}, K$) {

1) **INITIALISATION :**
   a) Choisir au hasard K centres de gravité : $G^{(0)} = \{g_1, \dots, g_k\}$
   b) Construire la partition de k classes : $C^{(0)} = \{c_1, \dots, c_k\}$
   où
   $c_j = \{ X \in Object / \forall\, i \neq j, d(x, g_j) < d(x, g_i)\ \forall j \in [1, k] \}$
2) **Boucle Principale :**
   a) Recalculer les centres : $G^{(t)} = \{g_1, \dots, g_k\}$
   où $g_j = \frac{\sum_{X \in j} X}{|c_j|}, \forall\, j \in [1, k]$
   b) Et les classes : $C^{(t)} = \{c_1, \dots, c_k\}$
   Test d'arrêt : $(C^{(t+1)} \cong C^{(t)})$.
3) **return** (la partition $C^{(t)} = \{c_1, \dots, c_k\}$)

}

- **RecTree**

RecTree est un algorithme de filtrage collaboratif appelé l'arbre de recommandation (Recommandation Tree). L'algorithme RecTree fractionne les données dans des cliques d'utilisateurs approximativement semblables. L'objectif est de maximiser les similarités entre les membres d'une même clique et à minimiser celles entre les membres de deux cliques déférentes.

Réseau Bayésien

Il est possible d'utiliser les réseaux de Bayes et les arbres de décision dans le contexte du filtrage collaboratif. En effet, un réseau de Bayes est un graphe acyclique dirigé qui représente une distribution de probabilités de dépendance entre un ensemble d'entités (utilisateurs ou ressources). Chaque nœud dans le graphe représente une entité, et chaque arc une dépendance directe entre variables. Ainsi, chaque variable est indépendante de ses non descendants dans le graphe, étant donné l'état de ses parents **[15]**.

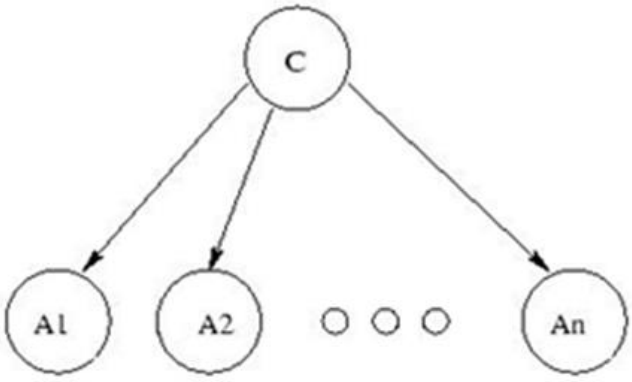

Figure 1. 8 - Réseaux de Bayes

Si on dispose d'un ensemble d'apprentissage, l'idée pour l'algorithme d'apprentissage est alors d'effectuer une recherche parmi les différentes structures de modèle possible, en termes de dépendance, pour que dans le réseau résultant, chaque ressource puisse avoir un ensemble de ressources parentes qui soient les meilleurs prédicateurs de ses votes.

Une idée est alors d'associer un réseau de Bayes à chaque ressource de la base. À chaque feuille de l'arbre est associée une probabilité d'attribuer une note à la ressource, étant donné l'état des parents identifiés. Pour prédire l'estimation de note d'un utilisateur sur une ressource, on se déplace alors dans le réseau de Bayes correspondant, selon les notes que l'utilisateur considéré a donné aux items parents présents dans le réseau, et on attribue, pour l'item considéré, la note la plus probable.

1.3.3. Filtrage hybride

Constatant les avantages et inconvénients de chacune des deux approches ci-dessus, on comprend que de nombreux systèmes reposent sur leur combinaison, ce qui en fait des systèmes de filtrage dits « hybrides ». En général, l'hybridation s'effectue en deux phases : (i) appliquer séparément le filtrage collaboratif et autres techniques de filtrage pour générer des recommandations candidates, et (ii) combiner ces ensembles de recommandations préliminaires selon certaines méthodes telles que la pondération, la mixtion, la cascade, la commutation, etc., afin de produire les recommandations finales pour les utilisateurs **[16]**.

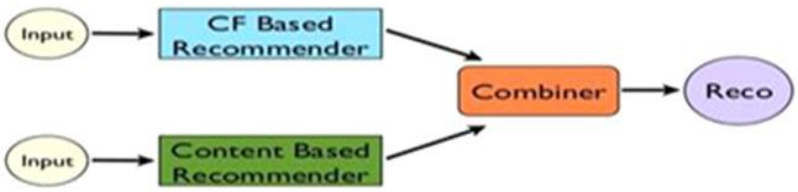

Figure 1. 9 - Le filtrage hybride.

Plus généralement, les systèmes hybrides gèrent des profils d'utilisateurs orientés contenu, et la comparaison entre ces profils donne lieu à la formation de communautés d'utilisateurs permettant le filtrage collaboratif. La meilleure description des méthodes hybrides a été faite par *(Burke, 2002)* **[2]**.

Alors, selon *Burke* on peut distinguer sept (07) façons de combiner les méthodes traditionnelles :

Pondération (Weighted)

Une méthode hybride qui combine la sortie d'approches distinctes, utilisant, par exemple, une combinaison linéaire des scores de chaque technique de recommandation.

Commutation (Switching)

C'est une technique qui permet de faire le choix d'un modèle de recommandation parmi plusieurs, en se basant sur plusieurs critères. La détermination de la technique appropriée dépend de la situation. Le système se doit alors de définir les critères de commutation, ou les cas où l'utilisation d'une autre technique est recommandée. Ceci permet au système de connaitre les points forts et les points faibles des techniques de recommandation qui le constituent.

Technique mixte (Mixed)

Dans cette approche, le recommandeur ne combine pas, mais augmente la description des ensembles de données, en prenant en considération les estimations des utilisateurs et la description des items. La nouvelle fonction de prédiction doit faire face aux deux types de descriptions et permet d'éviter les problèmes posés par le filtrage collaboratif, à savoir, le démarrage à froid.

Combinaison de caractéristiques (Features combination)

Dans un hybride basé sur la combinaison de caractéristiques, les données provenant de techniques collaboratives sont traitées comme une caractéristique, et une approche basée sur le contenu est utilisé sur ces données.

Cascade

La cascade implique un processus étape par étape. Dans ce cas, une technique de recommandation est appliquée en premier, produisant un ensemble de candidats potentiels.

Puis, une deuxième technique raffine les résultats obtenus dans la première étape. Cette méthode a pour avantage que si la première technique génère peu de recommandations, ou si ces recommandations sont ordonnées afin de permettre une sélection rapide, la deuxième technique ne sera plus utilisée.

Augmentation de caractéristiques (Feature augmentation)

L'augmentation de caractéristiques est semblable à la cascade, mais dans ce cas-là les résultats obtenus (le classement ou la classification) de la première technique sont utilisées par le deuxième comme une caractéristique ajoutée.

Méta niveau (Meta-level)

Dans un hybride basé sur méta niveau, une première technique est utilisée, mais differemment que la précédente méthode (augmentation de caractéristiques), non pas pour produire de nouvelles caractéristiques, mais pour produire un modèle. Et dans la deuxième étape, c'est le modèle entier qui servira d'entrée pour la deuxième technique *Arnautu* **[17]**.

1.4. Recommandation de ressources

Dans un réseau social, la navigation peut être effectuée à travers les ressources. Leur recommandation est une tâche importante pour l'adaptation de la navigation. Parmi les travaux qui adaptent la recommandation de ressources, nous trouvons ceux qui visent à recommander des items scientifiques **[18]** en combinant le CF et topic problistic modelling sans avoir recours à l'analyse les tags. Il y a des travaux qui recommandent des items (URLs) dans *Delicious* (*Wetzker 2008*) **[19]** ou encore des ressources (livres, articles, documents, images, audio, vidéo) **[20,21]**, en combinant le CF et le CB.

Dans le même contexte, (*Zheng 2011*) **[22]** proposent un système de recommandation de ressources basé sur l'historique et les connexions sociales (les relations d'amitié établies entre les utilisateurs) et aussi le temps des tags. Il utilise l'importance et l'utilité des tags afin de prédire les préférences de l'utilisateur et examine comment exploiter une telle information afin de construire un modèle de recommandation de ressource.

(*Brusilovsky 2010*) **[23]** proposent un système pour aider les étudiants a mieux avoir les ressources pertinentes, dans un contexte d'apprentissage en ligne (e-learning), en combinant le : i) Filtrage collaborative (CF) qui a pour but d'aider les utilisateurs a naviguer selon les informations de tous les utilisateurs, et ii) l'espace d'information enrichie par l'historique (History-enriched information spaces). Ce dernier fournit un support pour la navigation en rendant les actions de chaque utilisateur visibles pour les autres.

(*Beldjoudi, 2011*) **[24]** extraient des règles d'association a partir de la folksonomie (ensemble des tags) et les utilisent pour recommander des ressources. Cette méthode de recommandation réduit l'ambigüité des tags en prenant en considération les similarités calculées dans la folksonomie.

Ces travaux ne prennent pas en compte la sémantique des tags ce qui a effacé la qualité de la recommandation, vu l'ambigüité présente dans les tags.

(*Kim et al, 2012*) **[25]** utilisent la notation IEML[5] qui prend en compte la sémantique des tags afin de recommander des ressources (texte, image, vidéo) en se basant sur le

[5] www.ieml.org

comportement d'annotation. Cependant, cette notation est limitée car le vocabulaire utilisé est limité et n'englobe pas tous les mots.

Afin de recommander des ressources pertinentes basées sur le contexte, (*Joly et al, 2010*) **[26]** combinent les métadonnées (titre, description et contenu) avec la page visitée. Ils partent de l'hypothèse que le contexte de l'utilisateur peut être modélisé comme un ensemble de termes pondéré en combinant les métadonnées avec les tags. Ils combinent :

1. Les données extraites à partir des métadonnées,
2. Les données extraites à partir du tag,
3. Les données extraites à partir de l'analyse sémantique du contenu,
4. Les données extraites à partir de la localisation de la ressource,
5. Les données extraites à partir des annotations sociales.

La combinaison de ces critères permet d'avoir une précision des recommandations de 72%. Cette technique est intéressante surtout que les métadonnées décrivent le contenu réel de la ressource.

Aussi, (*Manzat et al, 2010*) **[27]** exploitent le comportement de l'utilisateur pour enrichir des métadonnées. Cet enrichissement est exploit pour l'adaptation de la présentation. Cet enrichissement servira pour faire des recommandations. Les métadonnées des ressources sont pondérées selon l'utilisation de l'utilisateur. Cette approche présente le concept de température.

La température reflète la popularité d'un document ou d'un élément de métadonnées a un moment donne. La température de métadonnées pour un certain groupe d'utilisateurs, a un certain moment, traduit l'intérêt de ce groupe pour la partie du document décrit par les métadonnées. Si la ressource n'est pas "consommée" dans une période de temps, le poids des métadonnées diminue. L'originalité de cette approche est que les métadonnées sont toujours gardées même si le poids est égal au zéro. Ceci est avantageux dans le cas de réapparition de la ressource, et donc le calcul du poids sera plus facile.

1.5. Recommandation de tags

Parmi les travaux de recommandation de tags basée CF, nous notons le célèbre système *Auto Tag* (*Mishne, 2006*) **[28]** qui suggère des tags pour les postes des weblogs. Plus tard, le système *Tag Assist* (*Sood et Hammond, 2007*) **[29]** est proposé pour améliorer l'approche *Auto Tag*. En effet, cette approche trouve les blogs postes similaires et suggère un sous ensemble de tags associes à travers le TSE (*Tag Suggestion Engine*). Ces systèmes présentent des inconvénients à savoir :

1) La formule de classification des tags est effectuée sur la somme des occurrences

de tags sur toute la folksonomie sans considérer la similarité avec la ressource annotée. De cette manière, les tags souvent utilisés pour annoter des ressources pas toujours similaires, peuvent être classes les premiers.

2) Le modèle proposé ne prend pas en considération l'ancien comportement d'annotation de l'utilisateur. Si deux utilisateurs annotent la même ressource, ils vont recevoir la même suggestion puisqu'une folksonomie construite à partir de ressources similaires est la même. Afin de répondre à ces inconvénients, (*Musto et al, 2009*) **[30]** introduisent *STaR*(*Social Tag Recommander System*), un système de recommandation de tags basé sur l'analyse des ressources similaires, en affectant un poids sur les tags déjà sélectionnés par l'utilisateur durant son ancien comportement d'annotation. Malgré les améliorations, ces techniques ne supportent pas une analyse sémantique sur les tags.

Les problèmes liés aux tags influencent la recommandation des tags et par conséquent influencent aussi la navigation. En fait, l'ambiguïté associée aux tags peut être un obstacle pour recommander des tags compréhensible pour les utilisateurs. De plus, ces tag doivent être relatifs à des intérêts et non pas des tags personnels (reflétant l'avis des autres utilisateurs).

1.6. Conclusion

Un système de recommandation peut aider l'utilisateur à trouver des informations en leur fournissant des suggestions personnalisées, cependant trouver des informations sur un grand site peut être un processus long et difficile.

Au sein de ce chapitre, nous avons présenté les concepts de bases relatifs aux systèmes de recommandation, en détaillant les différentes techniques utilisées par ces systèmes, à savoir la méthode du filtrage basé sur contenu et la méthode du filtrage collaboratif en expliquant le principe de chaque méthode. Ensuite, nous avons décrit les principaux avantages et inconvénients de chacune de ces méthodes. Toutes ces approches présentent néanmoins des caractéristiques complémentaires et par conséquent un grand nombre de travaux se sont intéressés aux approches hybrides qui combinent plusieurs approches et qui permettent de profiter des avantages et fournir des recommandations plus précises.

Enfin, nous avons terminé en citant les deux recommandations : celle des tags et celle des ressources.

Chapitre II

Systèmes d’annotations (tagging) collaboratives

Chapitre 2 : Systèmes d'annotations (tagging) collaboratives

2.1. Introduction

Marquage collaboratif, étiquetage collaboratif, Tagging social, annotations sociales ou Tagging collaboratif, différentes appellations désignant toutes ce phénomène qui est apparu ces dernières années et qui ne cesse de gagner une popularité sur le web 2.0. Marquer un contenu par des termes descriptifs est une manière d'organiser ce contenu pour une navigation future, un filtrage ou une recherche. Le Tagging collaboratif est devenu un moyen de plus en plus courant pour le partage et l'organisation du contenu web.

Du Tagging découle un autre concept qui est celui de la Folksonomie, un néologisme qui ne cesse de prendre de l'ampleur, et qui représente un système de classification et d'indexation des documents via des termes dans un esprit collaboratif.

Actuellement, de nombreux sites offrent la possibilité de tagguer du contenu. Divisés en catégories, il y a ceux spécialisés dans le partage des papiers scientifiques ou de références bibliographiques tels que *Connotea*, d'annotation de photos comme *Flickr*[6] ou de vidéos comme *Youtube* et *Dailymotion*[7] ou encore de signets (bookmarks) tel que *Delicious*.

Dans ce chapitre, nous allons définir le tagging collaboratif et la folksonomie. D'abord nous évoquerons les systèmes du tagging collaboratifs et leurs principales caractéristiques, leur conception ainsi que les différentes structures des actions du tagging. Enfin nous citerons quelques exemples de systèmes de tagging et nous terminerons par une présentation des limites de ces systèmes. Coté folksonomie nous citerons leurs types et caractéristiques, nous présenterons ce qu'elle apporte de plus par rapport à la classification traditionnelle, et enfin les futures perspectives dans ce domaine.

2.2. Systèmes de tagging

2.2.1. Définitions

Tag *:* Etiquette, libellé, descripteur, métadonnée ou même marqueur, c'est un terme choisi de façon informelle pour décrire une ressource dans le web, et qui peut prendre toutes les formes possibles et ça, selon la culture, l'expertise et la maitrise de l'utilisateur.

Selon *Guy Marieke* **[31]** :« *Essentiellement, un tag est simplement un jeu de mots-clés librement choisi. Cependant, du fait que les tags ne sont pas créés par des spécialistes de l'information, ils ne suivent aucune indication formelle*». Cela signifie que ces mots peuvent être catégorisés avec n'importe quel mot définissant une relation entre la ressource et un concept issu

[6] www.flickr.com
[7] www.dailymotion.com

de l'esprit de l'utilisateur. Un nombre infini de mots peut être choisi, dont quelques-uns sont issus de représentations évidentes tandis que d'autres ont peu de signification en dehors du contexte de l'auteur du tag.

De l'analyse d'une communauté de tagging dans le système *Del.icio.us, Golder* **[32,33]** avait déterminé qu'un tag peut avoir une multitude de fonctions. Parmi cette variété nous citons les sept (7) points suivants :

1. Identifier l'objet de quoi s'agit-il, son thème ou sujet : vacances, hiver.
2. En plus d'identifier le thème de l'objet (ou du contenu), un tag peut identifier ce qu'est l'objet lui-même : Une photo, un blog…
3. Identifier à qui l'objet appartient.
4. Description ou détail de tags existants: Certains tags n'ont aucune signification seuls. Ils n'ont de sens que quand ils sont associés à d'autres tags. Leur fonction est donc d'apporter plus de détails ou de description à des tags existants. Tel est souvent le cas avec les nombres comme par exemple le 10 de l'expression top 10.
5. Identifier des qualités ou des caractéristiques du contenu : C'est le fait de dire que tel contenu est 'comique', 'funny' ou 'horrible'…
6. Auto référence du tag : Dans ce cas, le tag illustre une relation entre l'utilisateur et le contenu. C'est le cas des tags commençant par 'mon' ou 'my' : myphoto, mon_enfant…
7. Aide-mémoire : Il s'agit de planifier une tâche donnée, 'à lire', 'à revoir'…etc.

Tag cloud : Le nuage de tag est une façon de présenter les tags correspondant à une ressource donnée ou l'ensemble des tags attribués par un utilisateur déterminé **[34].** La visibilité d'un tag dans le nuage augmente avec l'augmentation de son utilisation.

Tagging : On désigne par le tagging un processus qui consiste à créer des mots librement choisis -Tags- et de les associer à des objets (qui peuvent être : des signets, des photographies, de la musique ou des vidéos). C'est une organisation de ressource à travers l'étiquetage **[32].**

Tagging collaboratif : Indexation collaborative ou encore indexation sociale, c'est une pratique qui permet d'associer des mots-clés à une ressource. Une fois assignés, ces tags sont directement accessibles aux autres utilisateurs et exploitables dans le cadre de la recherche d'information. L'attribution de tags associe à une ressource une connaissance particulière ou un point de vue original de l'internaute et crée de multiples chemins d'accès à cette ressource.

Selon *Golder* **[35]** :« *Le Tagging Collaboratif décrit le processus par lequel plusieurs utilisateurs ajoutent des métadonnées à un contenu partagé* ».

2.2.2. Caractéristiques des systèmes de tagging

Par rapport aux autres systèmes de classification, le système de tagging offre plusieurs caractéristiques, parmi lesquels nous citons **[31,33]** :

- Possibilité de libeller le même objet de façon multiple.
- L'octroi de tags à une ressource lui offre de multiples chemins d'accès exploitables par la communauté internet.
- Contribution à la constitution de réseaux sociaux en permettant à l'utilisateur de trouver d'autres usagers partageant des centres d'intérêts communs.
- La certitude de répondre aux besoins des utilisateurs, vu que ces derniers taguent eux-mêmes les ressources.
- Les systèmes de tags fournissent un simple texte box pour entrer des balises textuelles.
- La sérendipité c.à.d. la faculté de trouver ce qu'on ne cherche pas. En d'autres termes, aider les utilisateurs qui n'ont pas d'idées fixes de ce qu'ils cherchent en utilisant la navigation à travers les tags (le tag Cloud).
- Les systèmes de tagging sont de nature inclusive et non-hiérarchique (un objet peut appartenir à plusieurs catégories).

2.2.3. Propriétés d'un système du Tagging

Un système de tagging collaboratif permet à ses utilisateurs d'attribuer des mots clés à des contenus partagés sur le web. Nous citons ci-dessous ce qu'offre un tel système pour l'utilisateur :

- Liberté du choix des mots clés (tags), en effet aucun contrôle n'est mis en œuvre pour guider ou forcer l'utilisateur excepté le recours aux ontologies et dans certains cas la suggestion de tags estimés adéquats par le système.
- Tagguer ses propres ressources, mais aussi les ressources créées et tagguées par d'autres utilisateurs.
- L'utilisateur doit être identifié, donc inscrit au préalable. Les informations requises changent d'un système à un autre. D'une manière générale, l'utilisateur est amené à introduire son nom, prénom, nom d'utilisateur, sexe et une petite description libre de l'utilisateur.
- L'utilisateur peut associer plusieurs tags pour la même ressource.
- Un même tag peut être associé à différentes ressources par différents utilisateurs.

- Une même ressource est tagguée par plusieurs utilisateurs ce qui donne l'aspect collaboratif.

2.2.4. Conception des systèmes de tagging

Il existe plusieurs facteurs qui contribuent à la conception des systèmes de tagging. Des études de *Golder et Huberman* **[32]** du système *Del.icio.us* et celle de *Marlow et al*., sur *Flickr* **[36]** , nous citons les attributs suivants:

- **Droits de tagging -Taggingrights-**: Une des dimensions les plus importantes lors de la conception d'un système de tagging. Un système peut être conçu en *self-tagging* où les utilisateurs ne taggue que leurs ressources (ex. : *Del.icio.us*), ou conçu en *free-for-all* où les utilisateurs peuvent tagguer toutes les ressources (ex. *Flickr*).
- **Support de tagging -Tagging support-** : Représente la visibilité de l'utilisateur, on distingue le *blind tagging* où l'utilisateur ne peut voir les tags assignés à une ressource (ex. *Dl.icio.us*), ou bien le *viewbale tagging* où l'utilisateur peut voir tous les tags assignés à la ressource (ex. : *CiteULike*[8]), ou le *suggetive tagging* où l'utilisateur ne peut voir que les tags suggérés par le système (ex. : *Yahoo*[9] *! MyWeb2.0*).
- **Agrégation des tags -Aggregation-** : Il existe plusieurs approches pour l'agrégation de tags, le *bag-model* où la multiplicité des tags pour la même ressource par différents utilisateurs est permise (ex. : *Del.icio.us*). Alternativement, le *set-model* où la répétition des tags est interdite (ex. : *YouTube*).
- **Type des ressources -Type of object-** : Elles peuvent être des images (ex. : *Flickr*), vidéos (ex. : *YouTube*), articles ou livres (ex. : *CiteULike*), pages web (ex. : *Del.icio.us*), musique (ex. :*Last.fm*),… (tout autre type d'objet qui peut être virtuellement représenté).
- **Liens entre ressources -Resource connectivity-** : Les ressources peuvent être dépendantes, directement liées (ex. : pages web), ou bien groupées (ex. : photos).
- **Liens entre utilisateurs –Social connectivity-** : Les utilisateurs peuvent être indépendants, liés ou bien groupés.

[8] www.citeulike.org
[9] www.yahoo.com

2.2.5. Motivations des utilisateurs

Les motivations de l'utilisateur affectent aussi les types de tags associés à une ressource donnée, ces motivations peuvent être :

- **Utilisation future** : Un utilisateur taggue une ressource pour se rappeler d'elle pour une utilisation ultérieure.
- **Contribution et partage** : L'utilisateur veut contribuer au Tagging d'une ressource au profit d'une ou de plusieurs autres personnes.
- **Attirer l'attention** : L'attention de l'utilisateur est attirée lorsque le système se présente dans un style visible et attirant (exemple les nuages de tags).
- **Jeux et compétition** : C'est le cas des jeux sur le Tagging tel qu'ESP Game.

2.2.6. Structure d'une action de tagging

La principale structure d'une action de Tagging est la structure tripartie, cependant il existe d'autres types de structure, à savoir la structure tripartie avec liens et la structure quadripartie.

Toute action de tagging est composée de trois éléments basiques : les utilisateurs, les ressources à tagguer et les tags utilisés.

- ***Utilisateurs*** : Acteurs responsables d'attribution de taggue à une ressource.
- ***Ressources*** : Elles peuvent être de divers types, c'est l'objet à tagguer.
- ***Tags*** : Termes attribués par un utilisateur pour décrire une ressource donnée.

2.2.6.1. Structure en tripartie

C'est la structure basique des actions de tagging, l'ensemble de ses instances est de forme (utilisateur, tag, ressource). La figure ci-dessous illustre cette structure.

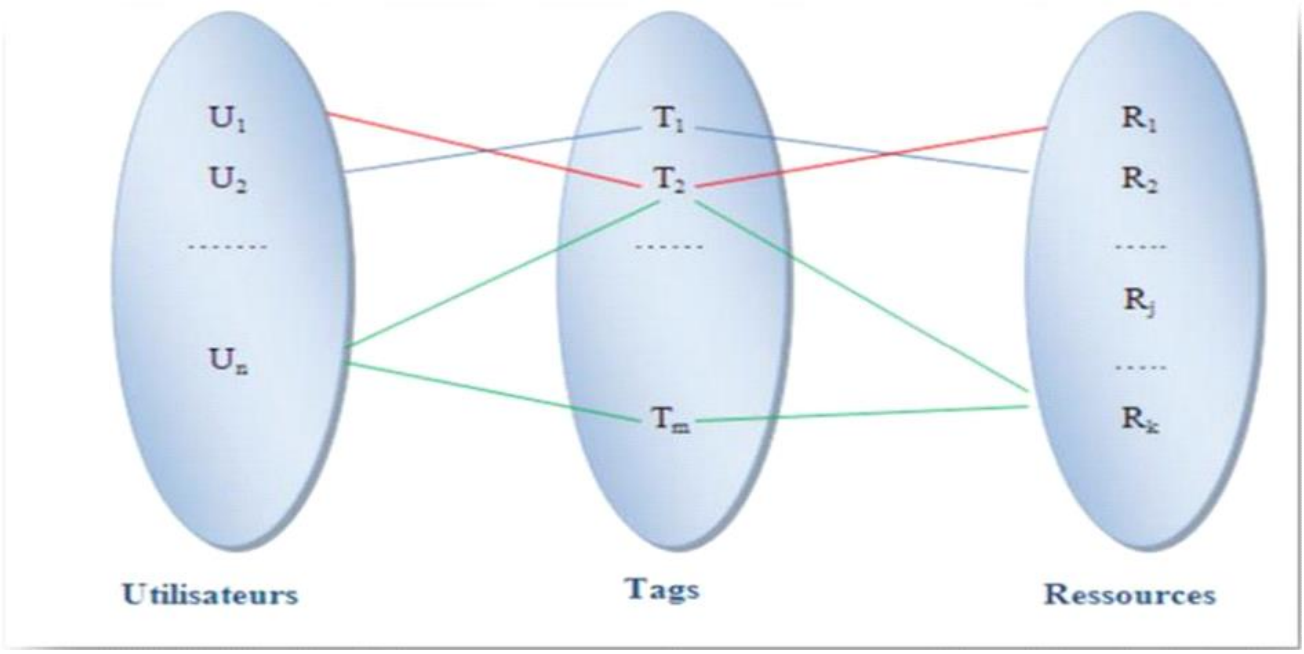

Figure 2. 1 - Action de tagging en tripartie [37]

Une instance de cette relation est composée d'un utilisateur, un ou plusieurs tags et une ressource : Tagging (utilisateur, ressource, tag).

2.2.6.2. Structure en tripartie avec lien inter ressources et inter utilisateurs

Basée sur le modèle de tagging en tripartie, cette structure est caractérisée par l'existence de liens entre les ressources (le cas des pages web) et des liens entre les utilisateurs (utilisateurs qui travaillent dans la même entreprise).

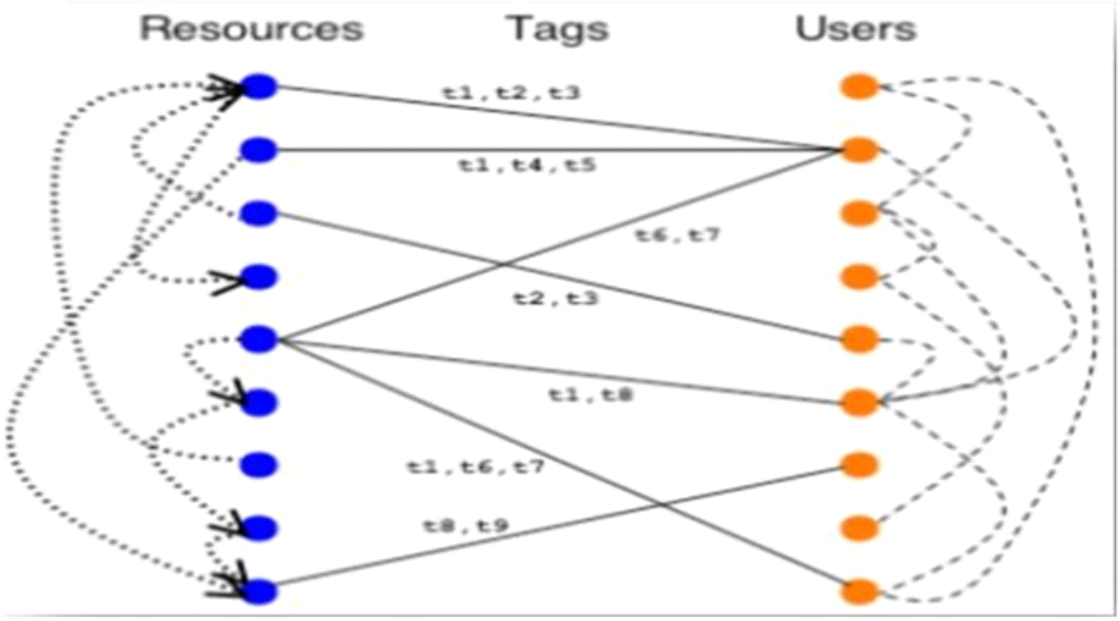

Figure 2. 2 - Action de tagging en tripartie avec liens inter-ressources et inter-utilisateurs [36]

Même si ces liens n'existent pas explicitement, on peut dire qu'il existe toujours des relations implicites entre ressources à travers les utilisateurs des tagguant (utilisateurs communs), et de même pour les utilisateurs à travers les ressources qu'ils tagguent (ressources communes entre un ensemble d'utilisateurs). On peut parler dans ce contexte d'émergence de relations.

2.2.6.3. Structure en quadripartie

La structure d'une action de tagging en quadripartie avait été proposé par *Gruber* **[38]**. L'idée de cette structure est d'ajouter la notion de sémantique à l'action de tagging. Dans ce cas-là, une action de tagging est représentée sous forme de quadruplet (utilisateur, ressource, tag, signification), où signification représente un concept de l'ontologie. La figure ci-dessous illustre le modèle proposé par Gruber.

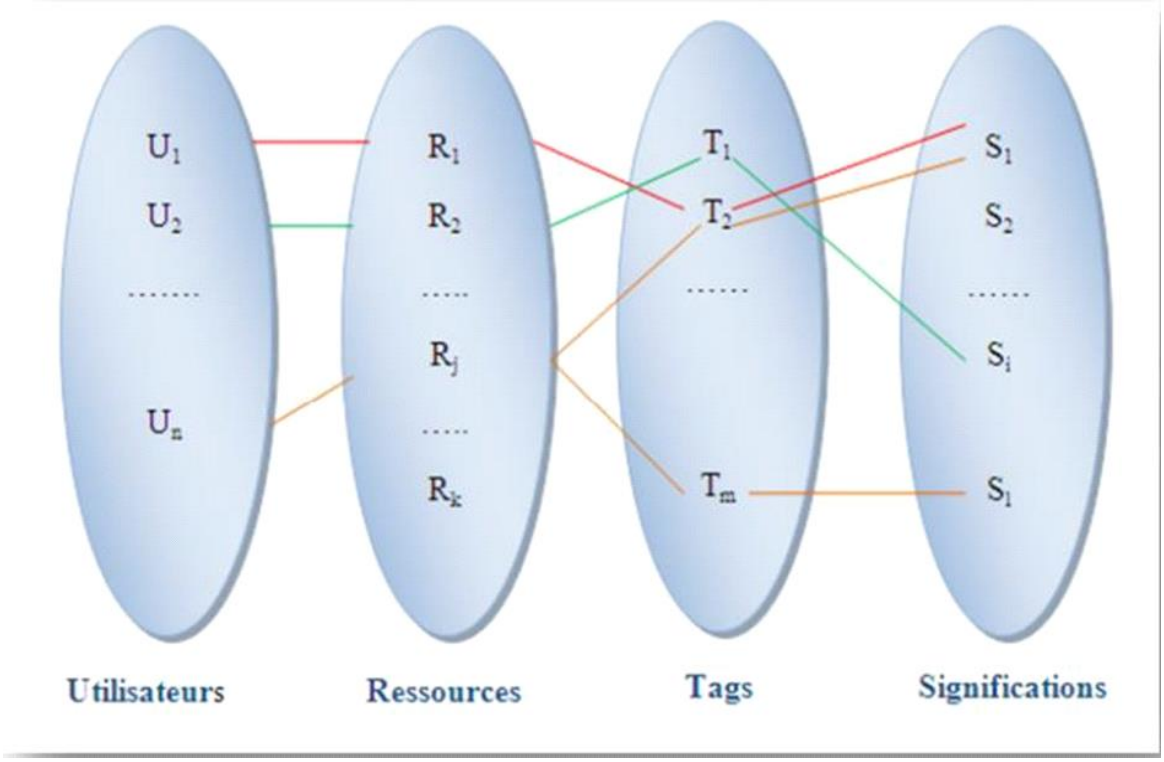

Figure 2. 3 - Action de tagging en quadripartie [38].

2.2.7. Quelques exemples de systèmes de tagging

- ***Flickr*** : « http://www.flickr.com» système de partage de photos qui permet de stocker et tagguer les photos personnelles ainsi que celles des autres contacts.
- ***YouTube*** : « http://www.youtube.com» système de partage de vidéos qui permet aux utilisateurs d'ajouter leurs vidéos et de les décrire avec des tags.
- ***Del.icio.us*** : « http://del.icio.us»-social bookmarking- ou gestion collaborative des signets, permet aux utilisateurs de sauvegarder et tagguer des pages web.
- ***Yahoo ! MyWeb2.0*** : « http://myweb.yahoo.com» similaire a Del.icio.us mais inclus un réseau social entre les contacts.
- ***CiteULike*** : « http://www.citeulike.org» site de tagging des citations et références académiques (Articles, livres,…).
- ***Last.fm*** :« http://www.last.fm»: Une base de données de musique, qui permet aux utilisateurs de tagguer des artistes, albums, chansons,…

2.2.8. Les limites des systèmes de tagging

L'application des systèmes de tagging à atteint beaucoup de domaine grâce à ses forces, néanmoins, ces systèmes présentent quelques limites, parmi lesquels nous citons :

- L'indexation par les utilisateurs semble moins couteuse, mais le temps passé à retrouver l'information s'accroit.
- Les systèmes de tagging peuvent être utilisés dans l'astroturfing, une pratique qui consiste à créer de faux profils utilisateurs par des entreprises pour faire croire aux internautes à la fiabilité de leurs produits **[31]**.
- Le problème de polysémie et synonymie demeure toujours **[32]**.
- Dans *Delicious*, l'un des systèmes les plus connu dans le Tagging, la valeur de la popularité d'un tag est le nombre de fois qu'il est cité pour tagguer l'ensemble des ressources. Le problème de variation d'écriture n'est pas résolu, exemple : Les tags *search_engine* et *searchengine* sont utilisés pour tagguer la page www.google.com et sont considérés par le système comme deux tags différents. Ces deux remarques sont valables pour pratiquement tous les autres systèmes. Un besoin en information de ressources tagguées avec *search_engine* n'inclura pas les ressources tagguées avec *searchengine*, (Silence).
- Dans tous les systèmes de Tagging, aucun critère n'est pris en considération pour différencier les utilisateurs du point de vue de leurs maitrise et expertise dans le domaine de la ressource. Par conséquent, deux tags associés à une ressource respectivement par un utilisateur expert et un novice auront le même poids.

2.3. Folksonomie

2.3.1. Définition

Le terme folksonomy, qui a donné en français folksonomie, a été inventé par un architecte de l'information, *Thomas Vander Wal*, qui, suite à de nombreuses demandes, a créé en février 2007 une page pour fixer sa version de la définition. Voici la définition du Grand dictionnaire terminologique de l'Office québécois de la langue française, datant de 2006 : « *Système de classification collaborative et spontanée de contenus Internet, basé sur l'attribution de mot-clés librement choisis par des utilisateurs non spécialisés, qui favorise le partage de ressources et permet d'améliorer la recherche d'information.* » La définition donnée par *T.Vander Wal* **[39,40]** permet de préciser la genèse du mot. L'auteur revient d'abord sur les origines des tags et du tagging, les systèmes de tags de la fin des années 1980 et des années 1990, leur arrivée sur le Web avec des services au début peu performants (Bitzi), en 2003 l'apparition de *del.icio.us*, premier service réellement coopératif permettant de voir les tags des autres, puis l'arrivée de *Flickr*, et les remous causé par cette nouvelle façon de trouver les choses, plus par sérendipité que par intention.

Les folksonomies sont des séries de métadonnées créées en collectif par les utilisateurs pour catégoriser et retrouver les ressources en ligne **[41]**. Une folksonomie est différente d'une taxonomie car, d'une part, elle n'est pas contrainte par des relations hiérarchiques, et d'autre part, elle n'est pas conçue par des experts. Il ne s'agit pas non plus d'une ontologie. Une ontologie est un ensemble structuré de concepts, alors qu'une folksonomie ne possède qu'une structure émergente, floue, et non contraignante (exemple un utilisateur peut utiliser un tag dans un sens totalement différent des autres utilisateurs).

La folksonomie est aussi connue sous l'appellation de potonomie, peuplonomie, taxinomie populaire ou même taxinomie sociale **[42]**.

Formellement, une folksonomie est un triplé F= <U, T, R>, où : **[43]**

- **U** : Représente l'ensemble des utilisateurs.
- **T** : L'ensemble des tags attribués par les utilisateurs.
- **R**: L'ensemble des ressources à tagguer.

Une folksonomie est un graphe G(F) = (N, A) avec :

- **N**: L'ensemble des nœuds. Ils sont représentés par les utilisateurs, les tags et les ressources.
- **A** : L'ensemble des arcs. IIs sont représentés par le triplet (u,t,r).

2.3.2. Caractéristiques

- Les folksonomies sont centrées sur les utilisateurs, ce n'est plus un seul professionnel qui indexe, mais plutôt toute une communauté d'usagers.
- Utilisation facile, multidirectionnelle et bidirectionnelle.
- Les folksonomies sont considérées à la fois comme des systèmes de description et des systèmes de classement **[44]**.
- Permet à l'internaute d'indexer des documents ou des informations et de les retrouver grâce à une classification des données à l'aide de mots-clés **[42]**.
- Evolutivité, où l'ajout d'un néologisme s'effectue très facilement et rapidement. En d'autre terme, les folksonomies permettent l'adoption rapide des nouveaux termes **[33]**.
- Les utilisateurs sont libres d'y choisir leurs propres mots-clés. La folksonomie est réellement centrée sur l'utilisateur. Il peut en faire une utilisation personnelle, professionnelle, etc. De cette façon, l'utilisateur peut devenir plus structuré et cela lui permet d'organiser toutes ses informations **[42]**.

- Les folksonomies ne reposent sur aucun thésaurus, ce qui confère à l'utilisateur une liberté totale quant aux choix des termes (tags) sans être contraint à une terminologie prédéfinie **[31,45]**.
- Les folksonomies sont très faciles à mettre en œuvre, aucune compétence n'est requise pour les produire **[42]**.

2.3.3. Type de folksonomies

Le concept de folksonomie recouvre deux types différents **[33]**. D'un côté, il y a les :

- ***Narrow folksonomies -Etroites- :*** Où seul le créateur de la ressource peut la tagguer, et donc absence de l'aspect collectif et communautaire des folksonomies. Ce type de folksonomie est connu aussi sous l'appellation de personomie. Un exemple très connu d'un système qui repose sur ce type de folksonomie est *Flickr*.

Et d'un autre coté il y a les:

- ***Broad folksonomies -Générales- :*** Où les ressources peuvent être tagguées par tout le monde. Un exemple très connu d'un système qui repose sur ce type de folksonomie est *Del.iciou.us*.

La deuxième différence majeure entre les deux types réside dans le fait qu'avec les narrows folksonomies, les tags ne peuvent apparaître qu'une seule fois par ressource, ce qui ne permet pas une bonne analyse des tags, contrairement aux broad folksonomies, qui permettent l'analyse de la fréquence des tags sur une ressource, et par conséquent l'obtention des tags les plus pertinents - les power tags-**[44]**.

2.3.4. Folksonomie Vs classification traditionnelle

Les folksonomies et les systèmes de classification traditionnelles (telle que la classification décimale de Deway DDC ou bien la classification de la Bibliothèque du Congrès LCC) représentent deux méthodes d'indexation d'objet via le web , ayant chacune des avantages et des inconvénients et se diffèrent sur plusieurs aspects. En se basant sur les travaux de *Mathe, A* **[46]** et *Zacklad* **[51]**, plusieurs points de comparaison sont tirés :

Folksonomie	Classification traditionnelle
- Indépendance dans l'utilisation des termes -tags-, et donc des mots Plus faciles à comprendre.	- Se base sur un vocabulaire contrôlé, et donc des mots non-familiers pour les utilisateurs.
- Basés sur l'opinion des utilisateurs simples.	- Nécessite l'implication de professionnels.
- Organisation non-hiérarchique, une ressource peut faire partie de plusieurs catégories en même temps.	- Organisation hiérarchique ce qui implique qu'un objet n'appartient qu'à une seule catégorie.
- Absence d'une standardisation de tags (fautes d'orthographes, pluriel,…)	- Standardisation des termes utilisés.
- Mise à jour fréquente, facile et immédiate.	- Mise à jour rare et complexe et nécessite un maintien de cohérence.
- Peu couteuse.	- Couteuse.

Tableau 2. 1 - Tableau comparatif des folksonomies et systèmes traditionnels d'indexation

2.3.5. Quelques règles pour une bonne indexation

Afin d'améliorer les systèmes de tagging, il serait bien intéressant de former les utilisateurs à l'indexation par tags. Inspirer des travaux d'**UlisesMejias [44,47]** qui consistait à établir une liste des tags inefficaces, des règles de bonne indexation par tag ont été édictées :

1. Penser d'une manière collective, car bien que les tags soient personnels, mais ils puissent être aussi utilisés par les autres.
2. Employer le pluriel pour définir des catégories.
3. Utiliser l'un des scores pour définir un groupe de mots.
4. Inclure les synonymes.
5. Eviter d'employer les majuscules.

Cet ensemble de consignes est très utile, notamment qu'une bonne indexation implique aussi le temps, car bien que l'indexation par les utilisateurs est moins couteuse, mais le temps passé à retrouver l'information peut s'accroitre à cause des tags inefficaces.

2.3.6. Avantages, limites et futures perspectives

Ces dernières années, les systèmes de tagging ainsi que les folksonomies qui en émergent se sont imposés au sein du web 2.0 comme le principal moyen de classification de données. La force du paradigme de folksonomie reste sur son utilisation de vocabulaire non contrôlé. De cette force résulte un nombre d'avantages. Nous citons **[31, 42,46]** :

- La souplesse d'utilisation des folksonomies.
- Adaptabilité, où l'insertion de nouveaux tags est une tâche très simple.
- Peu couteuse.
- Les folksonomies aident à accroitre et rendre plus rapide la diffusion de l'information.
- Mesurer la popularité d'une ressource, comme *Guten Tag* (site agrégateur de tags).

Certains caractéristiques des folksonomies leurs offrent plusieurs avantages par rapport aux systèmes de classification traditionnelles, néanmoins, ces mêmes avantages peuvent constitués une limite pour les folksonomies. *Adam MATHES* dans son item **[46]** avait jugé que : « *les folksonomies représentent en même temps ce qu'il a de meilleur et de pire dans l'organisation de l'information* ».

La faiblesse des folksonomies réside dans son utilisation de vocabulaire non contrôlé. Bien que ce point soit un des importantes caractéristiques des folksonomie, néanmoins, de son utilisation résulte quelques problèmes, parmi lesquels nous citons :

- La variabilité d'écriture des tags supposés équivalents, telles que électricité, électricite, electricite par exemple reste un problème d'hétérogénéité **[48]**.
- L'ambiguïté, où un tag peut avoir plusieurs sens pour des concepts différents **[46]**.
- L'utilisation des acronymes.

Plusieurs études sont menées afin de dépasser les limites et faiblesses des folksonomies. Quant au manque de structure, plusieurs auteurs comme *Heyman et al.,Hotho* **[33]** proposent de concevoir des algorithmes de conversion de folksonomies en taxonomies bien hiérarchisées. *Specia&Motta* **[49]** ont essayé de traiter le problème d'hétérogénéité des tags en proposant une solution qui consiste à mesurer la distance d'édition entre les tags, et à partir d'un certain seuil, deux tags peuvent être considérés comme équivalents. Une autre solution avait été proposée par *Cattuto et al.,* **[50]** qui se base sur le calcul de similarité entre tags en se basant sur les liens entre les tags, les ressources et les utilisateurs.

Quant à la deuxième faiblesse des folksonomie, le problème de contrôle de tags peut être résolu en utilisant la technique de jardinage « tag gardening » qui consiste à établir une révision, de façon manuelle ou automatique, des folksonomies **[32]**.

De cela, on remarque bien que l'avenir du tagging collaboratif serait dans la fusion des folksonomies avec le vocabulaire contrôlé, car chacun va apporter à l'autre ce qui fait sa force. L'hybridation entre des dispositifs professionnels de classification et les folksonomies semble bien prometteuse **[51]**.

2.4. Travaux sur le tagging et folksonomie

Le tagging social s'est récemment imposé dans le WEB 2.0 comme support à l'organisation et partages des ressources en permettant à l'utilisateur de classer ses ressources en leur attribuant des termes (tags). C'est un domaine de recherche qui avait incité plusieurs chercheurs. Dans ce qui suit, nous présentons quelques études de ces systèmes. Ces études tentent de traiter différents aspect liés aux systèmes de tagging :

2.4.1. Etudes sur la dynamique des systèmes de tagging

L'étude de la dynamique des systèmes de tagging est l'un des premiers axes de recherches. La première étude avait été proposée par *Golder et al.,* **[32]** où les auteurs avaient étudié le système *Del.icio.us.*

L'étude de la dynamique de ces systèmes est l'une des voies explorées par les auteurs. La dynamique de tels systèmes peut se mesurer en un ensemble de facteurs entre autres : les activités des utilisateurs (fréquences d'utilisation du système), variations du nombre de tags utilisé par utilisateur ou par ressource, les types de tags…etc. Une autre alternative de cette étude est celle proposée par *Halpin et al.,* **[37]**. L'objectif de ces deux études est de chercher une loi de distribution de la fréquence d'utilisation des tags. Selon *Golder*, après un certain seuil, la fréquence de chaque tag est fixée et donc la distribution des tags pour une ressource donnée devient stable au cours de temps. L'hypothèse d' *Halpins* était que les tags les plus utilisés pour annoter une ressource demeurent les mêmes, et que la distribution de leurs fréquences d'apparition suit une loi de puissance **[52]**.

Une autre étude proposée par *Marlow et Huberman* **[36]** traite les facteurs qui peuvent influencer les systèmes de tagging et leurs dynamiques. Dans cette étude, les chercheurs ont conclu qu'il existe une multitude de facteurs qui peuvent influencer la dynamique des systèmes de tagging. Ces facteurs peuvent être classés en deux catégories :

- ***Facteurs relatifs à la conception des systèmes de tagging :*** La différence de conception des systèmes de tagging a un impact signifiant sur les tags résultants de ces systèmes et leurs dynamiques. Chaque conception de système se base sur un ensemble d'attributs (déjà cités dans la section 2.4 de ce chapitre), les valeurs associées à ces attributs jouent un rôle primordial et ont un impact considérable sur la qualité et l'utilisation des tags générés par ce système. Dans le tableau suivant, nous résumons ces attributs ainsi que l'impact de leurs choix sur la conception et la qualité des systèmes de tagging :

ATTRIBUT	VALEUR	IMPACT
Tagging right	Self tagging Free-for-all	- Rôle du tag dans le système - Nature des tags résultant
Tagging support	Blind –Suggested Viewable	- Nature des tags résultant
Aggregation Model	Bag Set	- Poids des tags pour d'éventuelles statistiques
Object type	Textual Non Textual	- Nature des tags résultants
Ressource connectivity	None - Link - Groups	- Similarité des tags
Social connectivity	None - Link - Groups	- Nature des tags - Folksonomie localisée

Tableau 2. 2 - Attributs de conception des systèmes de tagging [36]

- ***Motivation des utilisateurs à tagguer :*** Les motivations de l'utilisateur à tagguer jouent un rôle très signifiant quant à la qualité des tags, et par conséquent, l'efficacité du système. Certains utilisateurs tagguent pour un besoin personnel, d'autre pour l'intérêt général, alors que d'autres ne prouvent aucun intérêt à cette pratique. Les utilisateurs ont tendance à tagguer pour un but organisationnel ou social, la première pratique afin de classer les ressources selon leurs besoins personnels, quant à la deuxième pratique, elle représente bien l'action de tagguer afin d'exprimer son opinion, ou bien les caractéristiques de l'objet taggué.

Généralement, les motivations qui incitent les utilisateurs à tagguer peuvent être résumées en ces points :

- ***Futures recherches*** : L'incitation à l'action de tagging est pour une raison personnelle afin de permettre un futur filtrage.
- ***Contribution et partage*** : Contribution à tagguer des ressources au profit d'autres personnes.
- ***Attirer l'attention*** : L'objectif du tagueur est d'attirer l'attention des utilisateurs à visualiser ses ressources.
- ***Présentation personnelle*** : Le but du tagueur est de laisser sa marque personnelle sur une ressource particulière.
- ***Expression d'opinion*** : L'objectif alors est de partager, via le tag, les jugements avec d'autres personnes.

2.4.2. Etudes pour rapprochement des folksonomies et ontologies

Le rapprochement des folksonomies et ontologie représente un axe de recherche très important. L'objectif principal de ces recherches est d'apporter une richesse supplémentaire et un support formel aux folksonomies afin de surmonter les problèmes liés à l'ambiguïté des tags et leurs impacts sur l'efficacité des systèmes de tagging.

Le premier travail que nous citons dans cet axe de recherche est celui de *Passant* **[53].** Dans son étude, il propose de structurer les folksonomies par l'intermédiaire d'un système centralisé qui est l'ontologie. Un tag va être une propriété d'un concept d'une ontologie contrôlée -*HasTag*-. Dans ce cas-là, les utilisateurs vont proposer des tags pour des concepts qui existent déjà. Encore, l'utilisateur peut soumettre à l'administrateur un concept et son tag si ce dernier n'existe pas. L'intérêt est alors de lever le problème d'ambiguïté des tags et enrichir l'ontologie avec les nouveaux concepts proposés par les utilisateurs.

Dans son étude, *Gruber* **[54]** propose le projet « TagOntology » pour construire une ontologie de folksonomie. L'idée est de construire une ontologie dédiée à la formalisation et conceptualisation de l'action de tagging. Dans ce cas-là, l'action de tagging va être structurée en quadruplet (utilisateur, tag, ressource, domaine).

Dans ce même axe, *Mika* **[55]** dans son étude propose de construire des ontologies légères à partir de l'analyse des folksonomies. Il se base sur le modèle en tripartie de l'action de tagging où les instances sont les ressources web attribuées par un utilisateur à une liste de concepts (tags). Ensuite, il utilise des méthodes d'analyse de réseaux sociaux pour construire des réseaux qui reflètent des liens entre concepts et en déduire les regroupements de termes et les relations de subsumption « is-a ». Cette tâche est établie de deux manières :

1. Utiliser le réseau liant les instances (ressources) aux concepts (tags) qui les caractérisent, et donc rapprocher les concepts qui ont plus d'instances en communs. Ce graphe concepts- instances permet de déduire les cooccurrences des tags.
2. Ou bien utiliser les réseaux liant les utilisateurs aux concepts et donc rapprocher les concepts qui ont plus d'utilisateurs en communs. Ce graphe concepts-utilisateurs permet de déduire les communautés d'intérêts.

2.4.3. Etudes pour exploitation des tags dans les systèmes de recommandation

Beaucoup de chercheurs dans le domaine de la recherche d'information se sont intéressés à exploiter les tags pour améliorer la recherche.

Dans leur étude, *Bischoff et al.,* **[56]** se sont intéressés à trois systèmes de tagging différents : *Del.icio.us* , *Flickr* et *Last.fm* pour étudier le comportement de l'utilisateurs vis-à-vis de la recherche avec tagging. Cela consiste à suivre leurs choix lors du tagging afin de déterminer s'ils utilisent les mêmes tags pour annoter ou rechercher une ressource. Les auteurs ont aussi étudié les tags en les classant d'abord en catégories (topic, time, location,…), ensuite voir la fréquence d'utilisation de chaque catégorie selon le type de la ressource.

Une autre étude menée par *Xu et al.*, **[57]** propose une recherche personnalisée à base du tagging. L'idée est d'utiliser le modèle vectoriel de la recherche d'information pour la comparaison entre le vecteur de la requête et le vecteur du contenu. En plus établir une autre comparaison entre le vecteur d'intérêt de l'utilisateur (qui n'est autre que l'ensemble de ses tags) avec le vecteur topic (qui est l'ensemble des tags associés à un contenu). Le résultat de la recherche va être la somme de ces deux comparaisons.

Enfin, une étude dans le domaine de la recherche d'information proposée par *Bao et al.*, **[58]** qui propose d'utiliser les tags pour le calcul de popularité d'un contenu. Les auteurs proposent un algorithme SPR (Social Page Rank) pour capturer la popularité d'un contenu de points de vue de ses utilisateurs.

2.5. Conclusion

Les systèmes de tagging collaboratifs décrivent le processus par lequel des utilisateurs ajoutent des métadonnées à des ressources. Ce paradigme avait émergé suite à l'avènement du web 2.0 et depuis, il suscite beaucoup d'intérêts et son utilisation ne cesse de s'accroitre grâces à l'ensemble de ses caractéristiques qui représentent ses points forts.

Dans ce chapitre, nous avons détaillé ce paradigme en citant ses caractéristiques, structure et conception. Nous avons mis l'accent aussi sur les folksonomies, leurs avantages et limites. Enfin nous avons présenté quelques études portées sur les systèmes de tagging et les folksonomies.

Chapitre III

Présentation des systèmes de tagging existants.

Chapitre 3 : Présentation des systèmes de tagging existants.

3.1. Introduction

L'avènement du web 2.0, centré utilisateur, a fait émerger une quantité très importante d'informations. Souvent partagées dans les médias sociaux, ces informations constituent un moyen pour guider les autres utilisateurs vers l'information recherchée. Cet aspect collaboratif de partage d'information est utilisé dans plusieurs applications comme le e-commerce, le e-learning, etc. Cependant, cette quantité d'informations rend leur accès de plus en plus difficile compte tenu de la diversité du contenu qui peut intéresser l'utilisateur. Plusieurs techniques ont été développées afin de mieux utiliser la connaissance collective partagée par les utilisateurs du réseau social. Parmi ces techniques, citons l'adaptation qui permet de fournir à l'utilisateur une information qui convient mieux à répondre à ses besoins. L'adaptation est un terme assez générique et se décline sous plusieurs formes : personnalisation, recommandation….etc.

L'adaptation est un processus fortement lié à l'utilisateur. En effet, nous adaptons l'information selon les besoins de chaque utilisateur. Donc, un profil utilisateur qui reflète les caractéristiques appropriées de l'utilisateur (intérêts, préférences….etc.) permet d'améliorer l'adaptation ainsi que d'éviter la surcharge cognitive et la désorientation de l'utilisateur pendant son accès à l'espace de l'information.

Dans un contexte social, l'utilisateur est de plus en plus actif (il participe aux discussions, commente et annote les ressources….etc.), mobile (il accède partout à l'information) et curieux (il compare pour avoir la meilleure information, cherche des avis ….etc.). Il a donc besoin d'informations adaptées reflétant ses besoins actuels et intérêts à chaque période de temps. Cela a pour but de lui fournir une meilleure adaptation lors de l'accès à l'espace d'information et pendant l'évolution de ses intérêts. En effet, les intérêts des utilisateurs peuvent changer et devenir non à jour dans le temps. Ainsi, un intérêt jugé pertinent dans une période de temps peut fluctuer dans la période suivante de temps.

3.2. Systèmes de recommandation à base de tags

Les systèmes de tagging offrent aux utilisateurs un autre moyen d'adresser les tâches de recommandation et de prévision. *Shirky* suggère que puisque les tagué sont crées par les utilisateurs, elles représentent des concepts qui leur sont significatifs **[77]**. Comme les tags sont facilement compréhensibles par les utilisateurs, les tags servent de passerelle pour permettre aux utilisateurs de mieux comprendre une relation inconnue entre un élément et eux-mêmes.

Les systèmes de tagging sociaux sont généralement utilisés pour faciliter l'indexation collaborative de quantité massive d'informations et améliorer leur accès **[59]**. Par exemple, les

services en ligne de gestion des références de bookmarking social *Connotea2* et *CiteULike3* sont utilisés par les chercheurs, scientifiques et académiciens pour stocker, organiser, partager et découvrir des liens vers des papiers académiques et de recherche. Le système ASK – LOST 2.0 **[60]** propose d'utiliser les tags pour indexer tous types de ressources pédagogiques digitales (images, vidéos, textes, URL). Les tags sont également utilisés comme outil d'indexation et de recherche d'informations dans des communautés d'enseignants, comme le propose le site Web *Couldworks*[10] **[61]** créé pour les enseignants afin de discuter de leurs pratiques et idées de design pédagogique. L'outil de partage de signets *SemanticScuttle* **[62]** propose également aux communautés de structurer leurs tags en créant des relations explicites d'inclusion et de synonymie entre tags.

Les systèmes de recommandation des médias sociaux sont un domaine de recherche jeune qui a récemment attiré une attention considérable, comme en témoigne le nombre croissant de publications (par exemple, **[63, 64]**) et se prépare à poursuivre sa croissance.

3.2.1. Détection des intérêts de l'utilisateur social

La création des medias sociaux a fait émerger de nouveaux comportements associés a l'utilisateur qui reflètent ses intérêts. En fait, l'utilisateur social n'appartient plus à l'audience mais devient un contributeur actif de la création du contenu social. Par conséquent, ses intérêts deviennent de plus en plus compliqués à détecter. Dans un contexte social, de nombreux paramètres font de la détection d'intérêts un problème complexe. Nous avons choisi de nous concentrer sur certains problèmes qui affectent le processus de détection des intérêts :

a) **Le manque d'information fournie par l'utilisateur lui-même :** L'utilisateur ne donne pas généralement de façon exhaustive les informations relatives à ses intérêts. Donc, le profil explicite de l'utilisateur ne peut jamais être considéré comme entièrement connu par le système. Ainsi, il est difficile de s'appuyer sur la seule analyse du profil explicite pour détecter les intérêts pertinents (*Tchuente et al., 2013*) **[65]**.

b) **L'activité dense de l'utilisateur social :** L'utilisateur est de plus en plus actif : il participe à des discussions, commente et annote des ressources…. etc. Par conséquent, détecter ses intérêts devient plus difficile (*Ma et al., 2011*) **[66]**. En effet, différents comportements de l'utilisateur peuvent décrire ses intérêts, ce qui rend ensuite le choix du comportement à analyser difficile.

c) **La variété et la quantité des ressources :** Une approche sociale de détection d'intérêts doit faire face à l'aspect évolutif des médias sociaux. La quantité d'informations (contenu

[10] www.cloudworks.co

et utilisateurs) est en croissance exponentielle. Pour les utilisateurs, de nombreuses relations peuvent être établies (relations d'amitié, utilisateurs appartenant au même groupe…. etc.). Pour le contenu, de nombreux types d'informations sont disponibles dans les médias sociaux tels que des images, pages web, vidéos….etc. Cette variété rend la détection d'intérêts plus difficile, puisque l'utilisateur peut interagir avec plusieurs contenus.

d) **La qualité potentiellement mauvaise des annotations (tags) :** En effet, les tags reflètent l'opinion d'un utilisateur vis-à-vis d'une ressource. Mais, ils sont générés par l'utilisateur (mots libres) et peuvent ainsi être non compréhensibles. Cette caractéristique peut induire une mauvaise compréhension des tags par le système ou même par les autres utilisateurs.

3.2.1.1. Le rôle du comportement social pour la détection des intérêts

L'utilisateur devenant un contributeur actif de création du contenu social adopte de nouveaux comportements, par exemple l'annotation des ressources avec des tags (comportement d'annotation). Ce comportement est souvent décrit comme la relation <Utilisateur, Tag, Ressource>. Des intérêts peuvent être extraits à partir de ce comportement d'annotation de l'utilisateur en analysant les tags utilisés **[67]**, les ressources **[66]** ou même les utilisateurs **[68]**.

D'autres recherches détectent les intérêts pertinents de l'utilisateur en combinant les intérêts détectés avec d'autres paramètres afin de créer des approches hybrides. Ces approches sont par exemple employées dans un contexte de recommandation comme **[69]** pour détecter les intérêts d'un utilisateur à partir d'autres utilisateurs qui ont visité les mêmes pages Web afin de recommander des pages Web. De même, **[70]** combinent les intérêts de l'utilisateur avec l'information extraite à partir de ses activités de tags, afin de recommander des tags. Et enfin, **[66]** combinent les intérêts de différentes sources avec un raisonnement sémantique, afin d'enrichir la liste d'intérêts.

Les intérêts de l'utilisateur sont des éléments clés pour aboutir à l'adaptation. Ils peuvent être extraits d'une manière implicite en observant l'interaction de l'utilisateur avec d'autres utilisateurs/ressources ou de manière explicite à partir du profil utilisateur directement.

Dans la littérature, les techniques de détection des intérêts de l'utilisateur ont été employés pour l'amélioration de la recherche d'information **[71]**, pour la recommandation de tags **[72]** et des ressources **[67]**….etc.

3.2.1.2. Approches de détection des intérêts à partir du comportement social

Selon **[73]**, les intérêts peuvent être déduits à partir de :

a) **L'utilisateur**, en détectant les intérêts en se basant sur ses liens sociaux. Par exemple les personnes qui communiquent beaucoup plus que d'autres, ont une connectivité plus élevée.
b) **L'objet**, en exploitant des intérêts communs basés sur l'accès des utilisateurs a des objets, par exemple : page Web, documents visites.
c) **Le tag**, en détectant des intérêts en analysant les annotations sociales (tags), par exemple les tags les plus récents, les tags les plus populaires, l'historique des annotations, ou même en analysant la sémantique des tags.

3.2.1.2.1. Travaux basés sur les utilisateurs

Dans un contexte social, les intérêts peuvent être déduits à partir d'autres personnes dans le réseau social (les personnes proches). La définition d'une personne proche est la relation sociale de l'utilisateur avec d'autres utilisateurs. Cette relation peut être explicite (une relation d'amitié) ou implicite (les utilisateurs qui agissent sur la même ressource par exemple).

Notons que la relation explicite se réfère à une relation de connaissance ou accointance entre personnes, alors que la relation implicite se réfère à une accointance d'action dans l'espace d'information. Les deux relations peuvent exister en même temps entre deux individus.

Ces relations sociales sont détaillées dans **[74]**. Les personnes proches de l'utilisateur dans un contexte social, sont décrites par des liens, ou un lien entre deux utilisateurs agrège tous les types de relations qui existent entre ces deux personnes **[74]**.

Les personnes proches sont considérées comme une source importante d'information. En effet, les informations issues des personnes proches ont prouvé leur utilité pour surmonter le problème de démarrage à froid (cold-start) pour les nouveaux utilisateurs du système (dans un contexte de réseau de citation et du réseau social *Facebook*[11]) **[75]**, pour détecter les intérêts des utilisateurs (dans DBLP et *Facebook*) **[65]**, **[76]** (dans un contexte de réseau social universitaire) et aussi pour enrichir les profils utilisateurs dans un but de recommandation **[67]** (dans *Delicious*) **[68]** (dans The Internet Movie DataBase (IMDB)).

Une personne proche peut être une "bonne" personne qui influence l'utilisateur d'une manière positive. Ceci servira pour dériver ou enrichir le profil de l'utilisateur selon l'information présenté dans les profils de ses personnes proches, pour recommander les

[11] https://www.facebook.com

ressources pertinentes selon les ressources visitées par ses personnes proches….etc. Si c'est une "bonne" personne, alors le résultat obtenu sera jugé profitable par l'utilisateur.

Une personne proche peut être aussi une "mauvaise" personne qui influence l'utilisateur d'une manière négative comme les "spammeurs" dont le but est d'inonder le système d'information visant à désorienter l'utilisateur. Si c'est une "mauvaise" personne alors le résultat obtenu sera jugé non profitable par l'utilisateur.

Des études analysent l'environnement social afin de détecter les personnes proches en termes d'intérêts :

- Ces personnes proches sont détectées par plusieurs métriques, telles que la similarité cosinus **[68, 22, 24]**, "X-compass" **[67]**, corrélation de Pearson **[21]**. Ces études rapprochent deux utilisateurs lorsque leurs intérêts sont considérés comme "proches".
- D'autres études détectent les personnes proches par des observations, comme **[68]**, qui enrichissent le profil d'un utilisateur avec les tags de ses personnes proches (ses amis) non inclus dans le profil. Ceci en se basant sur l'observation que deux personnes sont proches s'ils partagent des tags communs et donc, ils peuvent bien avoir des intérêts en communs. Egalement **[80]** supposent que deux utilisateurs sont proches s'ils partagent un grand nombre de tags, qui sont fortement connexés (de point de vue sémantique).
- D'autres chercheurs visent à combiner différents paramètres, afin de détecter la similarité entre les personnes proches. **[81]** calcule la similarité entre les auteurs scientifiques par une combinaison de leur proximité (le degré de séparation dans le graphe des co-auteurs), leur connectivité (le nombre des chemins dans le graphe entre deux auteurs) et le nombre de papier en commun. Dans le cadre de la recommandation sociale, **[82]** calculent le score de la proximité à partir de différents critères :
    a. Le nombre de personnes et/ou de tags dans le profil utilisateur qui sont liés à l'item.
    b. Le degré de connectivité de ces personnes et/ou des tags à l'utilisateur.
    c. Le degré de connectivité de ces personnes et/ou tags à l'item.
    d. La fraîcheur de l'item. (*Roth et al., 2010*) **[83]** détectent les relations implicites entre les utilisateurs par leur échange de courriers électriques. Ils calculent la proximité par la fréquence de l'interaction entre les utilisateurs, la fraîcheur de l'interaction (plus l'interaction est récente, plus elle est

importante) et la direction de l'interaction (les échanges de l'utilisateur vers les autres sont plus intéressantes que dans le cas inverse, afin d'éviter les spammeurs).

- Les personnes proches peuvent également être analysées dans un contexte basé sur les graphes, où (*Tchuente, 2013)* **[65]** analyse les réseaux egocentriques et les communautés pour dériver des profils utilisateurs (détecter les intérêts). Le processus de dérivation s'articule autour de quatre principales étapes successives, qui vont permettre de dériver les attributs de la dimension sociale du profil :
  a) Détection de communautés dans le réseau k-egocentrique.
  b) Profilage des communautés détectées,
  c) Caractérisation des communautés.
  d) Dérivation des attributs de la dimension sociale.

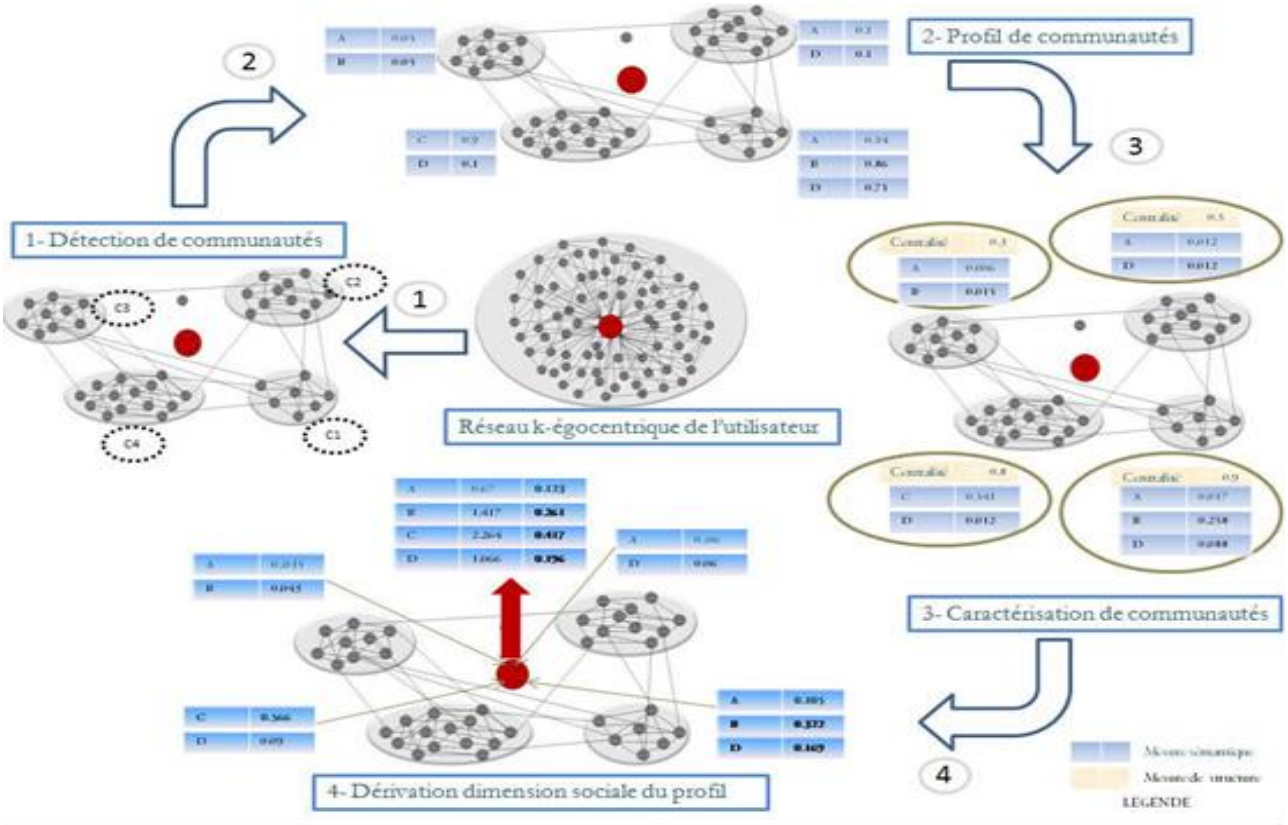

Figure 3. 1 - La dérivation du profil utilisateur selon les personnes [65].

A partir des travaux de (*Tchuente, 2013*) **[65]**, une étude à été menée dans **[84]** pour étudier l'influence de la base de données a dériver les intérêts de l'utilisateur à partir de son réseau egocentrique. Plus récemment, (*Rowe, 2014*) **[85]** détecte les intérêts des utilisateurs à travers un graphe de connexions/relations des utilisateurs. Les personnes proches d'un utilisateur sont définies comme ceux connectés à ce dernier.

3.2.1.2.1. Travaux basés sur les tags

Plusieurs recherches portent sur la détection des intérêts de l'utilisateur à partir de l'information sociale produite par les utilisateurs et particulièrement des tags. Ces derniers sont considérés comme une information puissante pour refléter l'opinion de l'utilisateur vis-à-vis d'une ressource **[67]** et aussi pour détecter les intérêts de l'utilisateur **[68]**.

Un tag peut être une façon de trouver des informations sur l'utilisateur selon son historique d'annotation **[85]**.

Les tags sont analysés de différentes manières. En fait, (*Nauerz et al., 2008*) **[86]** analysent le comportement d'annotation des utilisateurs pour détecter leurs intérêts. Ceci peut être utile pour recommander des informations ou faciliter l'accès aux informations.

(*Michlmayr, 2007*) **[87]** analyse les tags de l'utilisateur de différents manières : à travers une approche "naïve" qui agrège tous les tags, ii) une approche de co-occurrence entre les tags et en n iii) une approche adaptative. Leur objectif est de comparer les techniques et de trouver la plus performante dans un but de personnalisation de l'accès à l'information.

La détection des intérêts basés sur les tags, peut être effectuée en analysant les tags les plus récents **[88]**, les tags les plus populaires **[70, 22]**, l'historique des tags **[89, 90]**, les tags fournit directement par les utilisateurs **[104]** ou en analysant la sémantique des tags **[25]**.

Bien que les tags permettent d'obtenir les intérêts de l'utilisateur, ils sont des mots clés générés par l'utilisateur et donc ils ne suivent aucune règle précise. Par conséquent, ils peuvent contenir une information ambigüe et/ou qui ne reflète pas le contenu de la ressource. Par exemple un tag peut être : i) un spam (qui vise à promouvoir un intérêt d'un autre utilisateur par exemple) ou ii) un tag personnel (qui reflète le "sentiment" de l'utilisateur et non pas le contenu de la ressource comme : bien, j'aime, nul….etc.) ou bien iii) un mot propre à un utilisateur et qui n'est pas compréhensible soit par les autres utilisateurs soit par le système.

3.2.1.2.2. Travaux basés sur les ressources

Dans ces travaux, les intérêts des utilisateurs sont déduits sur la base des ressources que l'utilisateur accède **[66, 92]**. La ressource peut être de n'importe quel type (URL, vidéo, image, etc.). Dans **[66]** les intérêts des utilisateurs sont découverts par extraction et analyse des mots-clés de chaque source (les sources sont Facebook, *LinkedIn*[12]….etc.). Dans **[92]** les intérêts des utilisateurs sont découverts à partir de l'analyse du comportement de l'utilisateur à travers l'historique de visite des ressources, le temps passé sur une ressource.

[12] www.linkedin.com

Pour analyser le contenu de la ressource, différentes techniques existent telle que l'indexation qui est utilisée pour extraire les termes significatifs des ressources.

Après l'indexation des ressources, différentes fonctions de score peuvent être appliquées, afin de détecter la ressource la plus pertinente, selon une requête spécifique **[93]**. La précision d'une requête (par rapport à une ressource) peut être déduite par différentes fonctions de scores. Ces fonctions, appliquées dans un contexte de recherche d'informations, peuvent être TF-IDF, BM25….etc. **[93]**. Ces scores sont le résultat d'un calcul qui invoque une requête et une collection de ressources indexées. L'utilisation de ces méthodes a montré leur utilité et robustesse dans la recherche d'information **[95]**.

Dans un contexte social, la requête peut être un tag. Le contenu des ressources annotées a été analysé dans le but de recommandation dans une perspective d'apprentissage automatique dans **[72]**. En outre, (*Zhang et al., 2010*) **[94]** proposent une approche de recommandation (modélisée comme une approche basée sur Latent Dirichlet Allocation) dans les systèmes à base de tags. Cette approche combine le contenu et l'analyse des relations dans un modelé unique.

3.2.1.3. Approches d'enrichissement du profil utilisateur à partir du comportement social

Nous considérons un comportement social comme le comportement d'annotation constitué de l'information tag, ressource et utilisateur. Nous ne considérons pas dans notre étude l'analyse du comportement utilisateur sur les ressources (comportement de lecture, impression, clics….etc.). Nous détaillons les travaux d'enrichissement du profil utilisateur selon chaque type d'information.

3.2.1.3.1. Travaux basés sur les tags

Selon (*Meo et al., 2014*) **[96]**, l'utilisation de tags dénote implicitement les intérêts de l'utilisateur. De plus, analyser les tags de l'utilisateur est un puissant outil de gestion des connaissances.

(*Kim et al., 2011*) **[68]** enrichissent les intérêts des utilisateurs à partir du comportement d'annotation des utilisateurs pour recommander des ressources (dans The Internet Movie DataBase (IMDB)). Cette approche enrichit le profil utilisateur par des tags des voisins (personnes proches), en se basant sur l'hypothèse que l'utilisateur préfère les tags semblables publiés par ses voisins (voir figure). Ainsi, le processus d'enrichissement est fait selon les tags semblables des voisins non présents dans le profil utilisateur actuel. Cette approche a prouvé l'utilité de la connaissance collaborative (des voisins) pour améliorer la qualité des recommandations.

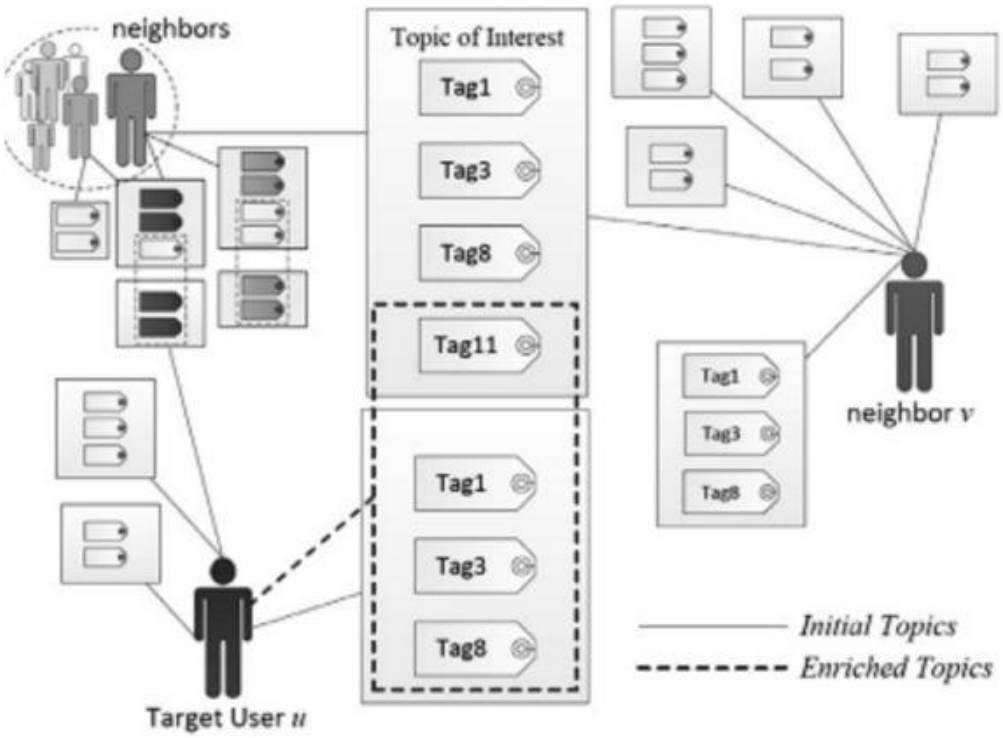

Figure 3. 2 - Enrichissement du profil utilisateur selon les tags des personnes proches [68]

(*Cantador et al., 2008*) **[97]** enrichissent le profil utilisateur par des tags issus de l'historique d'annotation pour améliorer les systèmes de recommandation. Cette approche associe les tags avec des ontologies pour incorporer le tag qui correspond au concept de l'ontologie. Cette approche utilise les différentes sources de l'historique de comportement d'annotation extrait de sites de médias sociaux populaires.

(*Meo et al., 2010*) **[67]** enrichissent les profils utilisateurs par les tags les plus "autoritaires" qui sont considérés comme les tags les plus importants (par exemple, les tags qui ont le PageRank le plus élevé). Cette approche considère deux graphes : TRG (Tag Resource Graph) et TUG (Tag User Graph). Ces graphes sont utilisés pour filtrer les tags qualitatifs par la méthode de **[98]**, générer la liste des tags candidats par la méthode de IDDFS (Iterative Deeping Depth First Search), et enfin fusionner ces listes de tags par la technique de Borda count. Finalement, le profil utilisateur est enrichi par les tags de ces listes. Cette méthode a montré que les tags sont automatiquement filtrés et classés en même temps par la technique Borda count. Cependant, elle ne considère pas la sémantique des tags, donc risque d'enrichir avec des tags ambigus et le contexte de l'utilisateur.

3.2.1.3.1. Travaux basés sur les ressources

L'évolution des documents numériques a mené à une classification de ces documents dans trois catégories : Les documents non structurés (document plat), les documents structurés (documents avec une structure explicite définie et connue à priori) et les documents semi-structurés (documents avec une structure flexible et un contenu hétérogène). Nous nous concentrons sur l'analyse des documents semi-structurés et plus précisément leurs métadonnées. Les métadonnées peuvent fournir des informations compréhensibles qui peuvent être utilisées pour l'interprétation des données **[99]**.

Les ressources sur les médias sociaux peuvent être considérées comme des informations puissantes qui reflètent les intérêts de l'utilisateur. En fait, les ressources annotées reflètent un intérêt potentiel, en tout cas non nul, de l'utilisateur par rapport à la ressource **[67]**. Par ailleurs, les ressources peuvent être notées (évaluées) et ceci reflète le degré d'intérêt de l'utilisateur vis-à-vis de ces ressources **[68]**.

Les ressources (semi-structurées) contiennent les métadonnées qui décrivent leur contenu.

Les métadonnées peuvent être utilisées dans un contexte d'adaptation comme la recommandation **[100, 101, 26]** ou d'enrichissement du profil utilisateur **[103]**. Nous détaillons chacune de ces recherches ci-dessous.

(*Bogers and van den Bosch, 2009*) **[100]** utilisent les métadonnées dans un but de recommandation. Cette approche utilise l'ensemble des tags (la folksonomie) et les métadonnées des articles pour stimuler la performance des algorithmes de filtrage collaboratifs traditionnels.

(Zitouni et al., 2012) **[101]** proposent une méthode pour la recommandation de ressources dans un contexte d'apprentissage en ligne. Ils utilisent les métadonnées pour recommander les nouvelles ressources. D'abord, ils extraient des métadonnées de nouvelles ressources. Ensuite, ils comparent la nouvelle ressource avec la collection préférée de l'utilisateur. En cas de similarité, ils envoient une notification à l'utilisateur.

(*Joly et al., 2010*) **[26]** proposent une méthode de filtrage et de recommandation de ressources. L'approche proposée agrège et interprète le contexte des données sur les terminaux des utilisateurs en forme de mots-clés pondérés (de tags). Ils calculent le poids du tag par rapport aux métadonnées de la page Web selon le nombre d'occurrences de chaque tag dans le titre, dans les mots-clés et dans la description de texte.

(*Abel et al., 2011b*) **[103]** exploitent des métadonnées (le titre, l'auteur, la date de publication du tweet) pour enrichir le profil dans le réseau social *Twitter*[13]. Les métadonnées sont

[13] www.twitter.com

utilisées pour connecter les tweets aux articles. Les tweets les plus liés sont utilisés pour enrichir le profil utilisateur.

Ces travaux, enrichissent le profil de manière qui ne prend pas le temps. L'enrichissement est effectué une fois pour toute.

3.3. Conclusion

Nous avons présenté dans ce chapitre, les systèmes de recommandation à base de tags où nous avons analysé les recherches portant sur la détection des intérêts de l'utilisateur social. Cette analyse utilise les informations du comportement d'annotation qui sont les tags, les ressources et les utilisateurs. Chacune de ces informations est importante pour refléter les intérêts des utilisateurs qui interagissent.

Chapitre IV

Proposition et évaluation.

Chapitre 4 : Proposition et évaluation

4.1. Introduction

Le marquage social (Tagging Social) joue un rôle de plus en plus important à la fois sur les plates-formes Web sociales tels que last.fm et *YouTube*, ainsi que sur les sites de commerce électronique à grande échelle tel qu'Amazon.com. Les applications Web sociales encouragent les utilisateurs à partager et à classer en collaboration le contenu à l'aide de tags.

À cet égard, nous pouvons dire que les recommandations ne sont pas évaluées simplement par leur valeur informative. Elles sont plutôt présentées dans un groupe informel d'utilisateurs et dans un contexte social, ce qui signifie qu'il existe une connaissance sociale particulière derrière les ressources recommandées. Bien que certaines activités de recherche récentes discutent de l'application des Tagging Social sur les systèmes de filtrage collaboratif, des approches plus avancées sont nécessaires pour mieux comprendre certaines activités telles que le marquage.

4.2. Notre système de recommandation basé sur les tags

Dans ce travail, nous visons à améliorer l'efficacité de la recommandation en incorporant les données sociales dans les algorithmes de recommandation traditionnels, et on essaye de montrer comment les tags peuvent être utilisés pour expliquer à l'utilisateur les recommandations générées automatiquement sous une forme claire et compréhensible de manière intuitive. A cet effet nous proposons une nouvelle approche de filtrage collaboratif basée sur les tags qui exploitent les données de marquage fournies par l'utilisateur pour recommander des items personnalisés aux utilisateurs.

Les systèmes de recommandation sociale traitent les problèmes de recommandation traditionnelle en exploitant les données sociales relatives à un utilisateur, à savoir son comportement de marquage, ses relations, son appartenance à des communautés, ses goûts, ses commentaires, ses votes, ses signets….etc. Ces données représentent implicitement les préférences concernant certains items ou des données contextuelles supplémentaires pour les médias enrichis. L'addition de cette information à la recommandation traditionnelle collaborative ou basée sur le contenu peut conduire à une amélioration significative.

Le marquage collaboratif peut aider les utilisateurs à organiser, partager et récupérer des informations de manière simple et rapide. Les informations de marquage collaboratif supposent des informations de préférences personnelles importantes pour l'utilisateur, elles peuvent être utilisées pour recommander des items personnalisés aux utilisateurs. Dans ce travail nous avons proposé une nouvelle approche de filtrage collaboratif basée sur les tags pour recommander des

items personnalisés aux utilisateurs. Cette dernière contient deux algorithmes basés utilisateur et basés item, où au sein de chacune nous avons utilisé trois matrices transformées à partir d'un espace tridimensionnel. Sur la base de relations tridimensionnelles distinctives entre les utilisateurs, les tags et les items, un nouveau procédé de mesure de similarité est proposé pour générer le voisinage des utilisateurs ayant un comportement de marquage similaire au lieu d'évaluations implicites similaires (respectivement items). La corrélation de Pearson et la similarité de cosinus **[21,22]** sont largement utilisés pour calculer la similarité en utilisant les données de notation explicites des utilisateurs. Cependant, les données de notation explicites ne sont pas toujours disponibles.

Dans l'algorithme basé sur l'utilisateur nous avons proposé une similarité globale basée sur le marquage et les activités sociales qui calculent la corrélation deux utilisateurs. La mesure de similarité proposée, inclut la combinaison de trois similarités partielles, à savoir : similarité basée sur des tags communs sur des items communs, sur les items et sur l'amitié. L'amitié est l'une des données les plus tangibles pour juger le comportement d'un utilisateur. Pour cette raison, nous avons décidé de quantifier la valeur de ces relations entre les paires d'utilisateurs afin d'évaluer leur proximité.

Pour l'algorithme basé sur l'item nous avons aussi proposé une similarité globale qui calcule la similitude entre deux items. Cette similarité est basée sur la combinaison de trois mesures de similarités, à savoir : similarité basée sur des tags communs entre les items, similarité basée sur des utilisateurs (items marqué par le même utilisateur) et similarité basée sur la relation utilisateur-tag. Le résultat expérimental prometteur montre qu'en utilisant les informations de marquage, l'approche proposée surpasse les approches de filtrage collaboratif standard basées sur les utilisateurs et les items. La sous-section suivante traitera de chaque approche en détail.

4.3. Mesurer la similarité

Il existe trois types d'algorithmes de filtrage collaboratif : basé sur la mémoire, basé sur le modèle et modèle hybride. Dans l'algorithme basé sur le modèle, les évaluations des utilisateurs sur les items sont collectées afin d'apprendre un modèle approprié. Ensuite, en utilisant le modèle construit, cet algorithme prédit les évaluations sur les items **[91]**. Construire un modèle adéquat n'est pas facile. Comparés aux algorithmes basés sur un modèle, les algorithmes basés sur la mémoire sont faciles à implémenter et plus pratiques. L'algorithme basé sur la mémoire recherche la similarité entre les utilisateurs ou entre les items en utilisant les évaluations des utilisateurs sur les items. En attendant, les voisins les plus proches d'un

utilisateur cible sont définis et, en fonction des voisins les plus proches, les items les plus intéressants sont recommandés à l'utilisateur.

Le but du calcul de similarité d'utilisateur est d'analyser la relation entre les utilisateurs. Si deux utilisateurs ont un profil similaire, il y a de fortes chances qu'ils agissent de la même manière à l'avenir. En outre, plus nous analysons le comportement d'un utilisateur en prenant en compte tous les aspects, plus nous comprenons les goûts de l'utilisateur et plus nous pouvons prédire avec précision ses préférences **[102]**.

Il existe différentes techniques que nous pouvons utiliser pour calculer la similarité entre les utilisateurs. Cependant, ces métriques ne se concentrent généralement que sur un ou deux facteurs **[102]**. Pour pouvoir mesurer correctement la similarité, une métrique de similarité doit refléter la compréhension par l'utilisateur de l'espace d'items de différentes perspectives. Dans ce travail, le calcul de similarité comprend deux algorithmes : l'un basé sur l'utilisateur et l'autre basé sur l'item.

Afin de faciliter la description de l'approche proposée, nous donnons d'abord les définitions suivantes :

U : Ensemble d'utilisateurs. $U = \{u_1, u_2 ... u_n\}$: Ensemble d'utilisateurs dans la communauté de marquage collaboratif.

P : Ensemble d'items. $P = \{p_1, p_2 ... p_m\}$

$P = \{p_1, p_2 ... p_m\}$: il contient tous les items taggués. Un item est un objet étiqueté par les utilisateurs et peut être n'importe quel type d'objet dans les domaines d'application, tels que les livres, les films, les URL, les photos, les documents académiques… etc.

T : Ensemble de tags. $T = \{t_1, t_2 ... t_l\}$: Ensemble de tags utilisés par les utilisateurs.

$E(u_i, t_j, p_\chi) = \{0,1\}$: Une fonction qui spécifie si l'utilisateur a utilisé le tag t_j pour taguer l'item p_χ

Profil utilisateur : Le profil de l'utilisateur sert à modéliser les fonctionnalités ou les préférences des utilisateurs. Les approches de profilage des utilisateurs avec une matrice d'évaluation d'utilisateur et des vecteurs de mots-clés sont largement utilisées dans les systèmes de recommandation. Pour profiler correctement et précisément le comportement des utilisateurs en matière de marquage, nous proposons de modéliser un utilisateur dans une communauté de marquage collaborative sous trois aspects : Les tags utilisés par l'utilisateur, les items marqués par l'utilisateur et la relation entre les tags et les items marqués.

Le profil utilisateur est défini comme suit:

Pour un utilisateur u_i, où $i = 1 ... n$.

Soit Tui le jeu de tag de ui, $\text{Tui} = \{\text{tj}|\text{tj} \in \text{T}, \exists \text{p}_\chi \in \text{P}, \text{E}(\text{u}_\text{i}, \text{t}_\text{j}, \text{p}_\chi) = 1\}, \text{Tui} \subseteq \text{T}$.

Soit Pui l'ensemble d'items de ui, $\text{Pui} = \{\text{p}_\chi|\text{p}_\chi \in \text{P}, \exists \text{t}_\text{j} \in \text{T}, \text{E}(\text{u}_\text{i}, \text{t}_\text{j}, \text{p}_\chi) = 1\}, \text{Pui} \subseteq \text{P}$.

Soit TPi la relation entre le tag et le l'item de l'utilisateur ui,

$\text{TPi} = \{\text{t}_\text{j} \in \text{T}, \text{p}_\chi \in \text{P}, \text{et } \text{E}(\text{u}_\text{i}, \text{t}_\text{j}, \text{p}_\chi) = 1\}$.

UFi = (Tui, Pui, TPi) est défini comme le profil utilisateur de l'utilisateur u_i. Le profil utilisateur ou le modèle utilisateur de tous les utilisateurs est désigné par UF, $\text{UF} = \{\text{UFi} \mid \text{i} = 1..\text{n}\}$.

4.4. Solution proposée :

Dans notre solution proposée, nous obtenons deux algorithmes de recommandation, basés sur les tags qui exploitent les données de marquage fournies par l'utilisateur. Dans ces deux algorithmes basé utilisateur et basé item, nous avons utilisé trois matrices transformées à partir d'espace tridimensionnel. Sur la base des relations tridimensionnelles distinctives entre les utilisateurs, les tags et les items, un nouveau procédé de mesure de similarité est proposé pour générer le voisinage des utilisateurs ayant un comportement de marquage similaire au lieu d'évaluations implicites similaires (respectivement items), la corrélation de Pearson et la similarité de cosinus **[22]** sont largement utilisées pour calculer la similarité en utilisant les données de notation explicites des utilisateurs. Cependant, les données de notation explicites ne sont pas toujours disponibles.

Dans le premier algorithme basé utilisateur nous proposons une nouvelle mesure de similarité d'utilisateur qui non seulement prend en compte les activités de marquage des utilisateurs, mais intègre également leurs relations sociales, telles que les amitiés, dans la mesure des voisins les plus proches. Dans le second algorithme basé item nous proposons une autre mesure de similarité d'item. L'étape suivante consiste à utiliser ce voisinage pour faire des recommandations et classer les items et recommander les N premiers items. Les résultats ont montré que l'introduction des annotations a été surtout bénéfique pour la recommandation personnalisée.

4.4.1. Algorithme basé sur l'utilisateur

Dans cet algorithme nous avons proposé une similarité globale, basée sur le marquage et les activités sociales qui calculent la corrélation entre deux utilisateurs. La mesure de similarité proposée inclut la combinaison de trois similarités partielles à savoir : similarité basée sur des tags communs sur des items communs, similarité basée sur les items et similarité basée sur l'amitié. L'amitié est l'une des données les plus tangibles pour juger le comportement d'un

utilisateur. Pour cette raison, nous avons décidé de quantifier la valeur de ces relations entre les paires d'utilisateurs afin d'évaluer leur proximité.

4.4.1.1. Similarité basée sur les tags (Tag-based similarity)

Plusieurs approches ont été proposées sur la manière d'améliorer une recommandation d'item dans les systèmes de marquage social. Cependant, ces approches considèrent soit des tags communs, soit des items communs entre utilisateurs. Mesurer la similarité des utilisateurs uniquement sur la base de tags communs n'est pas approprié, car il est possible que, même si deux utilisateurs possèdent de nombreux tags communs, la plupart de ces tags ne sont affectées à aucun item partagé par ces deux utilisateurs.

Notre première mesure de similarité utilise les informations de tag et nous nous concentrons uniquement sur les tags communs affectées aux items communs. Dans l'équation (4.1), nous définissons la valeur de similarité entre deux utilisateurs u et v.

$$\mathrm{UTsim}(u,v) = \frac{\sum_{i\in(I_u\cap I_v)}\left(\frac{(|T_{uv}|)^2}{|T_{ui}| * |T_{vi}|}\right)}{MAX(|I_u|,|I_v|)} \quad (4.1)$$

Où T_{ui} est un ensemble de tags que l'utilisateur u a attribué à l'élément i et T_{vi} est un ensemble de tags que l'utilisateur v a affecté à l'item i, T_{uvi} est un ensemble de tags communs, auquel les utilisateurs u et v ont été affectés à l'item i. I_u indique l'ensemble des items auxquels l'utilisateur a attribué des tags et I_v l'ensemble des items auxquels l'utilisateur v a attribué des tags. l'item i est dans l'intersection de I_u et I_v. Cela signifie que l'item i est un item commun entre l'utilisateur u et v. De plus, $MAX(|I_u|,|I_v|)$ indique le nombre maximal d'items sélectionnés par l'utilisateur u et le nombre d'items sélectionnés par l'utilisateur v. Dans cette équation, $\frac{(|T_{uvi}|)^2}{|T_{ui}|*|T_{vi}|}$ mesure en quoi l'opinion de l'utilisateur u et l'opinion de l'utilisateur v sont similaires sur l'item i (item commun entre u et v). Le chevauchement plus important entre les tags attribuées par les utilisateurs u et v à l'item i implique une plus grande similarité pour les opinions des utilisateurs u et v sur l'item i.

La raison d'utiliser $MAX(|I_u|,|I_v|)$ comme dénominateur est de normaliser notre métrique de similarité, car dans certaines situations, la somme de $\frac{(|T_{uvi}|)^2}{|T_{ui}|*|T_{vi}|}$ peut être supérieure à 1. Supposons que les utilisateurs u et v ont sélectionné n items. Si tous ces items sont identiques et que le jeu de tag de chaque item commun est identique, la valeur de

$\sum_{i\in(I_u\cap I_v)}\left(\frac{(|T_{uv}|)^2}{|T_{ui}|*|T_{vi}|}\right)$ est n. Par ailleurs, à des fins de comparaison, il convient de normaliser la similarité de deux utilisateurs pour obtenir une valeur inférieure ou égale à 1. Afin de déterminer un dénominateur approprié, nous analysons trois candidats pour le dénominateur:

- Le nombre d'items sélectionnés par l'un des utilisateurs :

  $|I_u|$ or $|I_v|$: Si dénominateur est le nombre d'items sélectionnés par l'utilisateur u ou le nombre d'items sélectionnés par l'utilisateur v, $\mathrm{Tsim}(u,v)$ n'est pas égal à $\mathrm{Tsim}(v,u)$ et la similarité n'est pas symétrique.
- Le nombre d'items communs sélectionnés par deux utilisateurs :

  $|I_u| \cap |I_v|$: Si le dénominateur est le nombre d'items communs entre les deux utilisateurs, un problème important peut survenir. Si l'utilisateur u et l'utilisateur v ne partagent aucun élément commun, le dénominateur est 0.
- Le nombre maximum d'items sélectionnés par deux utilisateurs :

  MAX $(|I_u|, |I_v|)$ Dans ce cas, les valeurs de similarité sont des valeurs plus raisonnables.

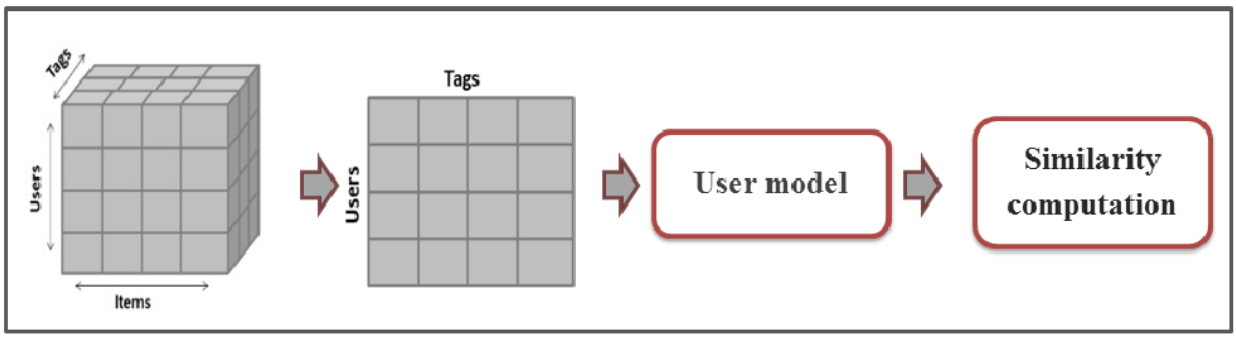

Figure 4. 1 - Similarité basée sur les tags

4.4.1.2. Similarité basée sur les items (Item-based similarity)

UP$sim(u,v)$: La similarité des items de l'utilisateur, qui est mesurée par le pourcentage d'items communs tagués par les deux utilisateurs:

$$\mathrm{UP}sim(u,v) = \frac{|\mathrm{Pu} \cap \mathrm{Pv}|}{\max\{|\mathrm{Pu}|, |\mathrm{Pv}|\}} \quad u,v \in U \qquad (4.2)$$

Où UP$sim(u,v)$ est la similarité entre les utilisateurs u et v, $|\mathrm{Pu} \cap \mathrm{Pv}|$ est un ensemble d'items communs entre l'utilisateur u et l'utilisateur v, Pu est un ensemble des items que l'utilisateur u tagué et Pv est un ensemble des items que l'utilisateur v tagué, $\max\{|\mathrm{Pu}|, |\mathrm{Pv}|\}$

indique le nombre maximal d'items sélectionnés par l'utilisateur u et le nombre d'items sélectionnés par l'utilisateur v, $Pu = \{p_\chi | p_\chi \in P, \exists t_j \in T, E(u_i, t_j, p_\chi) = 1\}$.

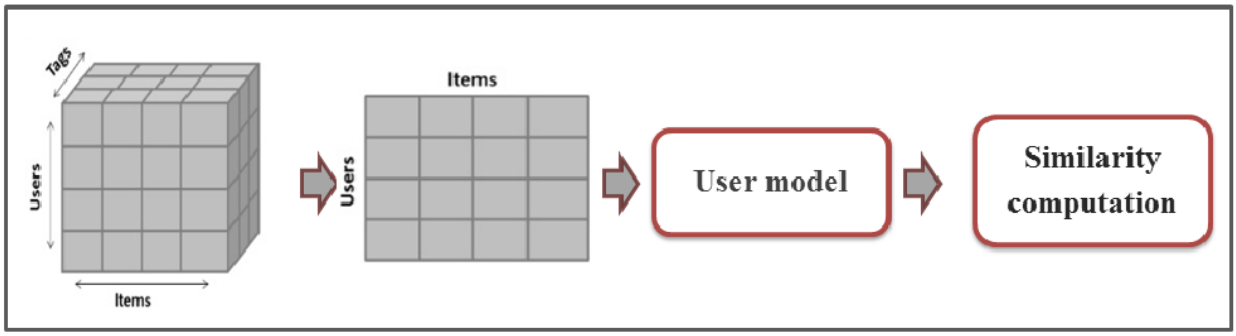

Figure 4. 2 - Similarité basée sur les items

4.4.1.3. Similarité basée sur l'amitié (Friendship-based similarity)

Un réseau social joue un rôle important en reflétant les intérêts partagés entre les entités dans un système basé sur le social. Pour faire face au problème de surcharge d'information dans le sens d'une énorme quantité de choix dans la sélection des items, un utilisateur peut faire confiance aux opinions de ses amis afin de filtrer les informations non pertinentes. Récemment, un nombre croissant de développements sur les sites web de réseaux sociaux fournissent une abondance de données sociales disponibles. De nombreux chercheurs ont étudié l'intégration des informations de réseau social avec le filtrage collaboratif basé sur le voisinage afin d'accroître la précision des systèmes de recommandation. La relation d'amitié est la plus populaire sur les sites web de réseaux sociaux **[105]**. Les relations sociales atténuent non seulement certaines limitations des relations implicites (telles que le problème de fragmentation des données **[106]**), mais peuvent être aussi potentiellement appliquées pour renforcer l'intelligence collaborative dans la recherche d'utilisateurs clés qui ont réellement un impact sur la prise de décision des autres.

Dans cette section, nous proposons une nouvelle mesure de similarité pour combiner les relations implicites et explicites afin d'accroître l'efficacité de la recommandation. Il a été prouvé que la combinaison de l'étiquetage social et l'amitié peuvent améliorer les performances d'un système de recommandation d'article **[107]**. Une relation d'amitié a un effet significatif sur la similarité des utilisateurs. Même dans ce cas, nous ne pouvons pas dire que la similarité d'utilisateurs devrait être simplement amplifiée s'il existe une relation d'amitié entre eux. La raison en est que deux utilisateurs peuvent être amis dans un système de marquage social, mais ils ne peuvent partager aucun intérêt commun sur la plupart des éléments. Par conséquent, nous devons rechercher les amis en qui l'utilisateur a fait confiance principalement et qui ont partagé les mêmes intérêts avec l'utilisateur **[21]**.

La figure ci-dessus présente le pseudo-code permettant de calculer la similarité en fonction des informations d'amitié.

```
1   Float  SimFriendShip (User u , User v)
2    {
3        if (v is friend of u)
4          {
5              try:
6                if (row.Tcom==0 ){
7                        FSimuv = row.UPsim ;
8                 }elif (row.UPsim==0){
9                       FSimuv = row.Tcom ;
10                }else{
11                     FSimuv = (row.Tcom + row.UPsim)/2 ;
12                }
13           }else
14               FSimuv = 0 ;
15               return FSimuv ;
16              except:
17                   print('erreur') ;
```

Figure 4. 3- Pseudo-code permettant de calculer la similarité d'utilisateur sur la base d'informations d'amitié

La fonction de similarité dans la ligne (1) de l'algorithme 1 accepte deux utilisateurs u et v en tant qu'entrées, génère un score de similarité basé sur l'amitié pour les utilisateurs qui remplissent des conditions particulières comme spécifié dans la ligne (3). Fondamentalement en ligne (3), nous vérifions si les utilisateurs u et v sont des amis. Si cette condition est vraie, alors dans la ligne (6), nous vérifions si UTsim $= 0$ est vraie, puis dans la ligne (7), un UPsim sera attribué à FS$imuv$. Si UP$sim = 0$ est vrai, à la ligne (9), la similarité UTsim sera attribué à FS$imuv$ et si UTsim! $= 0$! et UPsim ! $= 0$, nous utilisons l'équation de la ligne (11) pour calculer Fs$imuv$. Sinon, à la ligne (14), un '0' sera attribué à FS$imuv$.

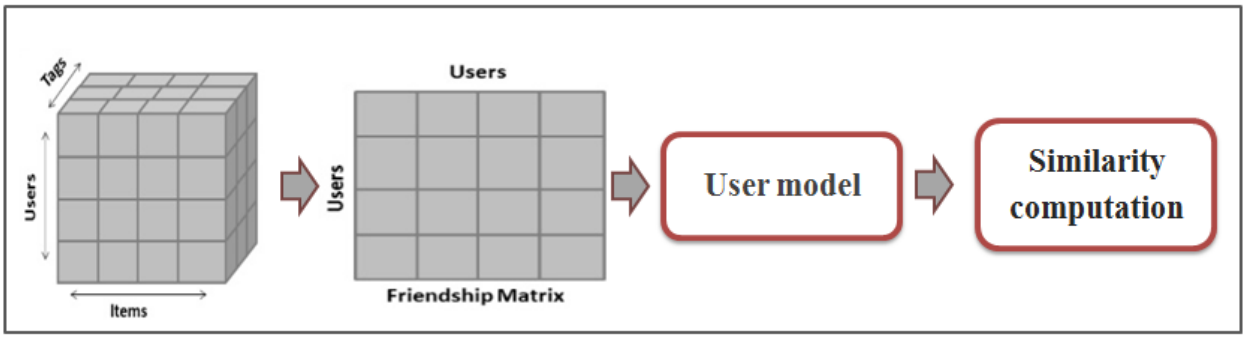

Figure 4. 4 - Similarité basée sur l'amitié.

4.4.1.4. Similarité globale

Afin d'améliorer l'exactitude des recommandations, tout d'abord, UPsim est calculé, puis les informations de similarité basées sur des tags sont incorporées. Ensuite, la méthode d'amplification est appliquée afin de prendre en compte l'amitié pour ceux qui entretiennent une relation étroite avec un utilisateur cible. Dans le but de calculer la similarité globale, nous avons défini deux paramètres, α et β, qui permettent d'ajuster le poids de différents facteurs.

$$\mathrm{Simu}(u, v) = \alpha * \mathrm{UPsim}(u, v) + (1 - \alpha) * (\beta * \mathrm{UTsim}(u, v) + (1 - \beta) * \mathrm{FSimuv}(u, v)) \quad (4.3)$$

Dans l'équation (4.3), la valeur précise de α et β doit être déterminée de manière empirique. Pour conserver la valeur de similarité globale entre 0 et 1, considérons 0 <α, β <1. Dans l'équation (4.3), α est appliqué pour ajuster le poids entre la similarité basée sur les items $\mathrm{UPsim}(u, v)$ et les deux autres similarités($\mathrm{UTsim}(u, v)$, $\mathrm{FSimuv}(u, v)$). Ensuite, β ajuste les poids relatifs entre ces deux similarités qui sont la similarité basée sur l'amitié $\mathrm{FSimuv}(u, v_j)$ et la similarité basée sur le tag $\mathrm{UTsim}(u, v)$. En ce sens, plus β est grand, plus le poids de l'activité de marquage sera important. Ceci fait, l'activité de marquage joue un rôle plus important. D'autre part, une valeur α plus grande implique que la similarité basée sur l'item joue un rôle plus important dans la similarité globale, en ajustant ces deux valeurs, nous déterminons quel facteur joue un rôle plus important dans notre décision de calculer la valeur de similarité.

Après le calcul de $\mathrm{Simu}(u, v)$ pour trouver des voisins, l'étape suivante consiste à recommander des items aux utilisateurs en prédisant chaque élément.

4.4.1.5. Génération de recommandations

L'une des étapes les plus importantes des systèmes de recommandation consiste à prédire le comportement futur d'un utilisateur. Dans un premier temps, un sous-ensemble d'utilisateurs similaires à un utilisateur cible, en fonction de leurs similarités puis l'agrégation pondérée de leurs évaluations est appliquée afin de formuler des recommandations à l'utilisateur **[108]**.

Étant donné que les tags sont des principaux attributs attachés aux items, pour les utiliser nous proposons une méthode pour formuler des recommandations d'items à l'utilisateur cible u, à savoir une approche basée sur l'utilisateur.

Pour l'approche basée sur les utilisateurs, le score de prédiction est calculé par la formule (4.4) en utilisant les similarités d'utilisateurs.

$$A^{u}(u_i, p_\chi) = \frac{\sum_{v \in C(u)} simu(u_i, v) * R(v, p_\chi)}{|C(u_i)|} \quad (4.4)$$

Soit u_i un utilisateur cible et $C(u)$ le voisin de u_i. Pour l'approche basée sur l'utilisateur, les items candidats pour u_i sont extraits des items étiquetés par les utilisateurs dans $C(u_i)$. Pour chaque item candidat p_χ, basé sur la similarité entre u_i et ses utilisateurs voisins, un score de prédiction noté $A^u(u_i, p_\chi)$ est calculé à l'aide de l'équation (4.4). Selon les scores de prédiction, les tops N items seront recommandés à u_i.

Les notations implicites des utilisateurs voisins en p_χ, notées $R(v, p_\chi)$, si un utilisateur a marqué un produit ou un item, la note implicite de cet item est définie sur 1 sinon 0.

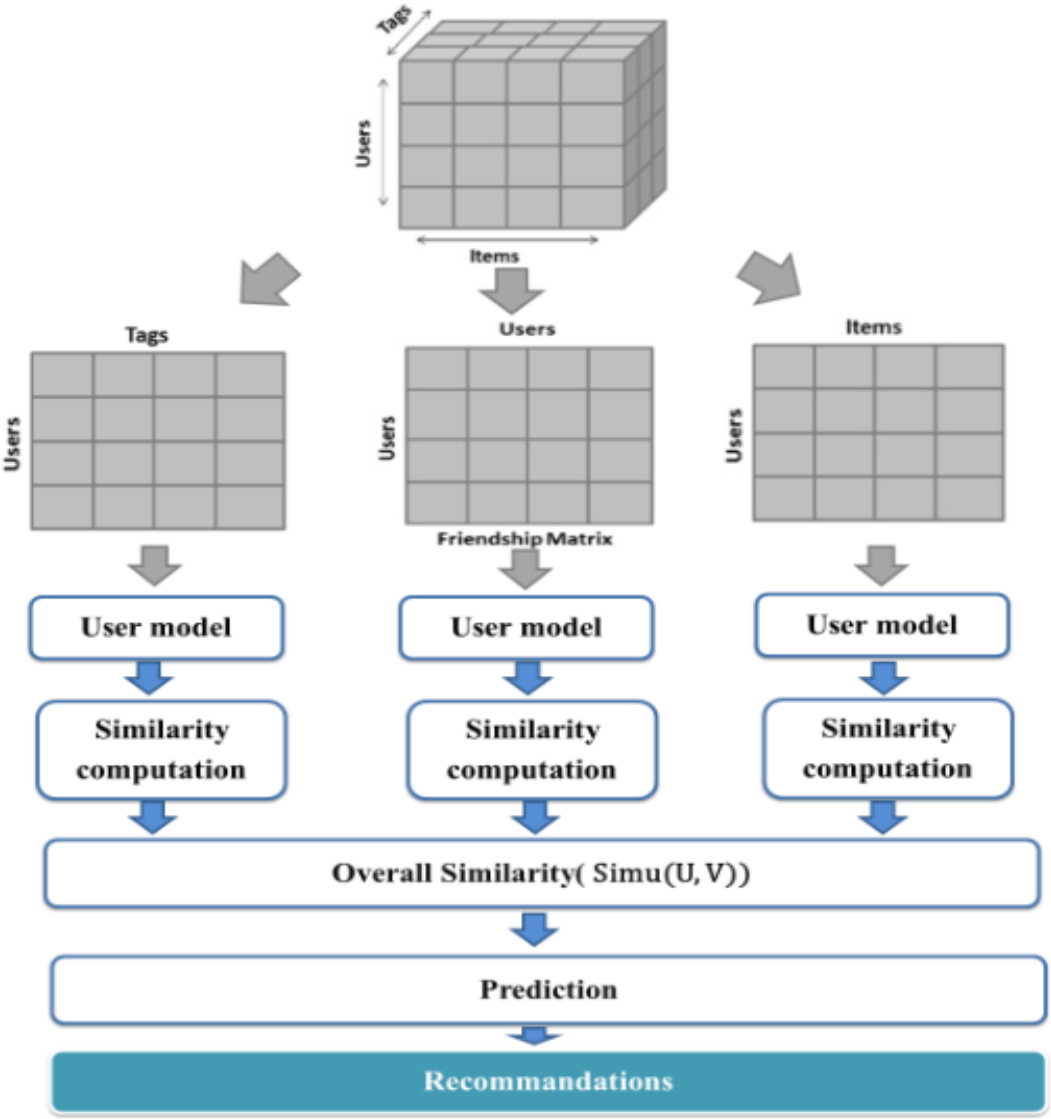

Figure 4. 5 - Approche basée sur l'utilisateur

4.4.2. Algorithme basée sur l'item

L'approche basée sur l'item se compose de deux étapes. La première étape consiste à calculer la similarité entre les items. La mesure de similarité doit non seulement inclure le nombre de tags communs aux items, le nombre d'utilisateurs ayant tagué le même item, mais

également le nombre d'utilisateurs utilisant le même tag. La deuxième étape consiste à faire des recommandations aux utilisateurs. L'ensemble d'items proposé peut être tous les items, à l'exception des items déjà notés ou étiquetés par l'utilisateur cible. Pour éviter le calcul inutile de paires d'items, les K premiers items les plus similaires de chaque item noté ou étiqueté de l'utilisateur cible u peuvent être agrégés ensemble en tant qu'ensemble d'items candidats.

4.4.2.1. Similarité basée sur utilisateur-tag relation (user-tag Relationship similarity)

PUT$sim(p_i, p_j)$: La similarité entre deux items basés sur commun utilisateur-tag Relationship.

$$\text{PUT}sim(p_i, p_j) = \frac{|\text{UT}_i \cap \text{UT}_j|}{\max\{|\text{UT}_i|, |\text{UT}_j|\}} \quad \text{P}_i, \text{P}_j \in \text{P} \qquad (4.5)$$

Où PUT$sim(p_i, p_j)$ est la similarité entre les items p_i et p_j, UT_i est l'ensemble de la relation utilisateur-tag de l'item p_i, UP_j est l'ensemble de la relation utilisateur-tag de l'item p_j, $\max\{|UT_i|, |UT_j|\}$ indique le nombre maximal des relations utilisateur-tag de l'item p_i et le nombre maximal des relations utilisateur-tag de l'item p_j.

$UT_j = \{< u_i, t_j > \mid u_i \in U, t_j \in T, \text{et } E(u_i, t_j, p_\chi) = 1\}$.

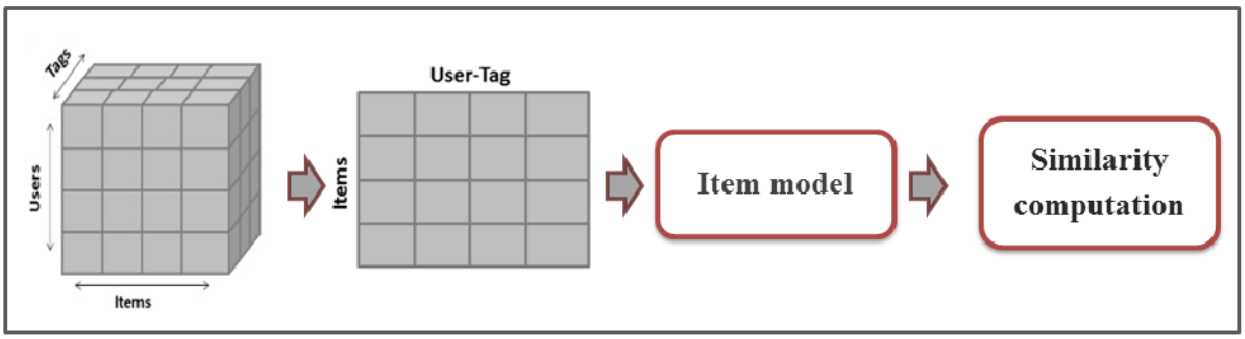

Figure 4. 6 - Similarité basée utilisateur-tag relation

4.4.2.2. Similarité basée sur les utilisateurs (User-based similarity)

PUS$sim(p_i, p_j)$: La similarité de deux items en fonction du pourcentage d'être tagué par le même utilisateur :

$$\text{PUS}sim(p_i, p_j) = \frac{|\text{Up}_i \cap \text{UP}_j|}{\max\{|\text{Up}_i|, |\text{UP}_j|\}} \quad \text{P}_i, \text{P}_j \in \text{P} \qquad (4.6)$$

Où PUS$sim(p_i, p_j)$ est la similarité entre les items p_i et p_j, Up_i est un ensemble des utilisateurs que sélectionne l'item p_i, UP_j est un ensemble des utilisateurs que sélectionne l'item

P_j, $\max\{|Up_i|, |UP_j|\}$ indique le nombre maximal d'utilisateurs qui sélectionne l'item p_i et le nombre d'utilisateurs qui sélectionne l'item p_j,

$Up_\chi = \{u_i | u_i \in U, \exists t_j \in T, E(u_i, t_j, p_\chi) = 1\}$.

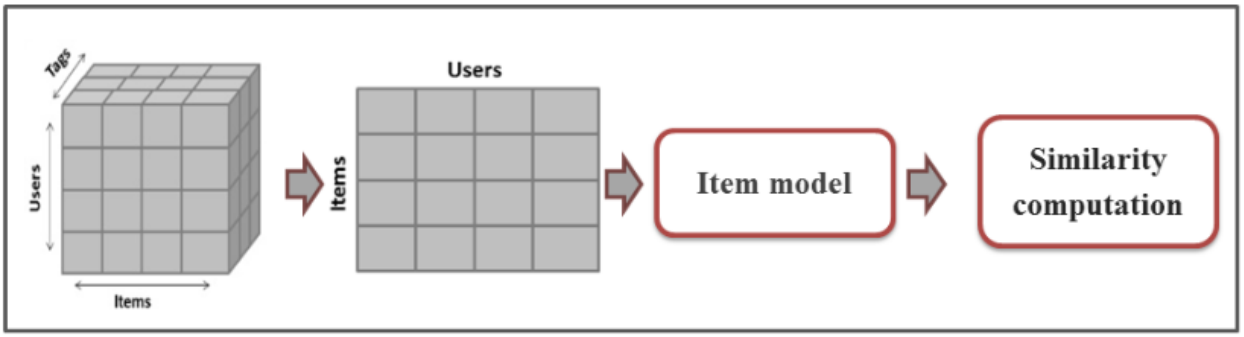

Figure 4. 7 - Similarité basée les utilisateurs

4.4.2.3. Similarité basée sur les tags (tag based Similarity)

Afin de trouver la similarité des items en considérant leurs fréquences de marquage, nous avons utilisé la méthode de similarité pondérée de Jaccard **[79]** avec quelques modifications. La méthode de similarité Jaccard originale ne prend pas en compte la fréquence de marquage et nous apportons donc quelques modifications. Selon l'équation (4.7), il existe un vecteur v_i pour chaque item Pi et un vecteur v_j pour l'item Pj, dans lequel chaque item de chacun de ces vecteurs est une paire de (Tag, Fréquence) représentant le tag attribuée à cet item et son contenu. La fréquence. Ainsi, $v_i(t).Fq$ détermine la fréquence de le tag t (un tag commun entre les items Pi et Pj) de l'item Pi. De même, $v_j(t).Fq$ détermine la fréquence du tag t sur l'item Pj.

$\text{Min}(v_i(t).Fq, v_j(t).Fq)$ renvoie le minimum de deux valeurs de fréquence de le tag t dans le vecteur v_i et le vecteur v_j. De plus, $\text{Max}(v_i(t).Fq, v_j(t).Fq)$ renvoie le maximum de deux valeurs de fréquence du tag t dans le vecteur v_i et le vecteur v_j.

$$PTsim(p_i, p_j) = \frac{\sum_{t \in (v_i \cap v_j)} \text{Min}(v_i(t).Fq, v_j(t).Fq)}{\sum_{t \in (v_i \cap v_j)} \text{Max}(v_i(t).Fq, v_j(t).Fq) + \sum_{ta \in (v_i \cup v_j - v_i \cap v_j)} \text{Max}(v_i(ta).Fq, v_j(ta).Fq)} \quad (4.7)$$

Dans l'équation (4.7), la numération correspond à la somme des fréquences minimales, soit la somme des valeurs de fréquence minimales de deux tags communes à l'item P_i et à l'item P_j. Au dénominateur, pour toutes les tags communes entre les items P_i et P_j, nous calculons la somme des valeurs de fréquence maximales de ces tags communes. De plus, au dénominateur de l'équation (4.7), ta est un tag qui n'est pas partagée entre les items P_i et P_j. $\text{Max}(v_i(ta).Fq, v_j(ta).Fq)$ renvoie le maximum de deux valeurs de fréquence du tag ta dans le vecteur v_i et le vecteur v_j. Enfin, nous trouvons la somme des valeurs de fréquence des tags qui

ne sont pas courantes entre les items P_i et P_j et ajoutons cette valeur à la somme des fréquences maximales pour obtenir le dénominateur.

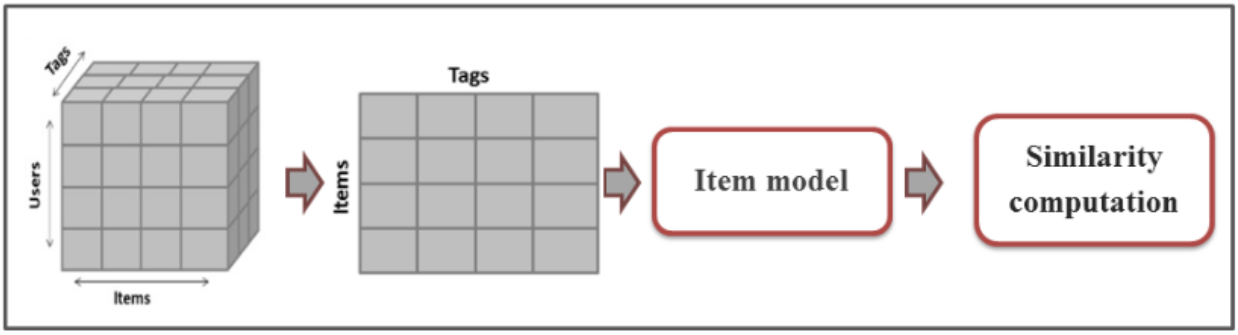

Figure 4. 8 - Similarité basée sur les tags

4.4.2.4. Similarité globale

La mesure de similarité globale de deux items est définie comme suit:

$$\mathrm{Simp}(p_i, p_j) = \alpha * \mathrm{PUTsim}(p_i, p_j) + (1-\alpha) * (\beta * \mathrm{PUsim}(p_i, p_j) + (1-\beta) * \mathrm{PTsim}(p_i, p_j)) \quad (4.8)$$

Dans l'équation (4.8), la valeur précise de α et β doit être déterminée de manière empirique. Pour conserver la valeur de similarité globale entre 0 et 1, considérons 0 <α, β <1, α est appliqué pour ajuster le poids entre la similarité basée sur utilisateur-tag Relationship ($\mathrm{PUTsim}(p_i, p_j)$) et les deux autres similarités ($\mathrm{PUsim}(p_i, p_j)$, $\mathrm{PTsim}(p_i, p_j)$). Ensuite, β ajuste les poids relatifs entre ces deux similarités qui sont la similarité basée sur l'utilisateurs $\mathrm{PUsim}(p_i, p_j)$ et la similarité basée sur le tags $\mathrm{PTsim}(p_i, p_j)$.

4.4.2.5. Génération de recommandations

Le score de prédiction peut être déterminé en calculant la somme de la similarité de l'item candidat avec tous les items évalués. Ainsi, si un item candidat a le score de similarité le plus élevé par rapport à l'un des items marqués de l'utilisateur, et qu'il présente les sujets les plus similaires préférences de sujet de l'utilisateur, cet item aura un score de prédiction plus élevé que les autres items. Nous proposons donc de calculer le score de prédiction d'un item candidat sur la base de la somme de la similarité avec chaque item étiqueté / évalué et de la similarité avec les préférences de sujet de l'utilisateur cible, comme indiqué ci-dessous.

$$Ap(u_i, p_\chi) = \sum_{p_j \in pu_i} simp(p_\chi, p_j) \qquad (4.9)$$

Où :

Ap (ui, p_χ) : Score de prédiction

simp(p_χ, p_j) : Est la valeur de similarité de l'item pk et de l'item p_j.

Pu_i : Est un ensemble des items de l'utilisateur u_i

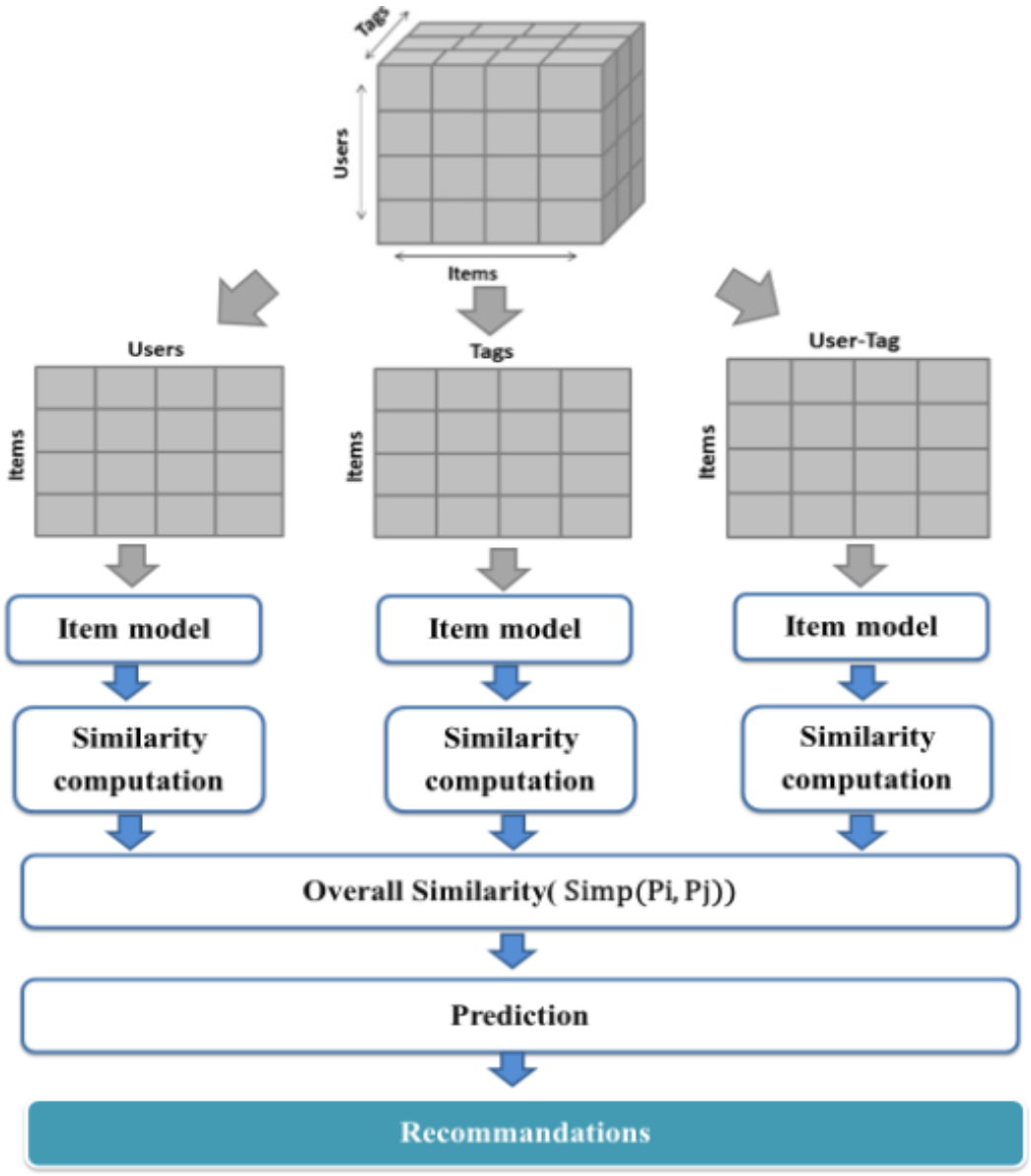

Figure 4. 9 - Approche basée sur l'item

4.5. Dataset

a) Description du jeu de données

Nous utilisons pour l'expérimentation, le jeu de données réelles du système de recommandation d'artiste musical. Ces jeux de données ont été publiés par la conférence HetRec 2011 et sont disponibles gratuitement à des fins de recherche et sur le site Web de la conférence

(http://ir.ii.uam.es/hetrec2011/datasets.html). L'ensemble de données est publié dans le cadre du deuxième atelier international sur l'hétérogénéité et la fusion de l'information dans les systèmes de recommandation (HetRec 2011). Dans HetRec 2011, il y a plusieurs jeux de données avec différents types de préférences utilisateur concernant les ressources appartenant à trois domaines (films, pages Web et pistes de musique) et contenant diverses méta-informations.

Nous utilisons dans cette expérimentation le jeu contenant 11946 tags, fournies par 1892 utilisateurs sur 17632 artistes et 186479 assignations de tags (tas), c'est-à-dire des tuples [utilisateur, tag, artiste], 98.562 tas par utilisateur et 14.891 tas par artiste.

Cet ensemble de données a été obtenu à partir du système de musique en ligne Last.fm. Ses utilisateurs sont interconnectés dans un réseau social généré à partir de relations "amis" Last.fm. Chaque utilisateur possède une liste des artistes de musique les plus écoutés, des attributions de balises, c'est-à-dire des tuples [utilisateur, tag, artiste], et des relations amicales au sein du réseau social de jeux de données. Chaque artiste a une URL Last.fm et une URL d'image.

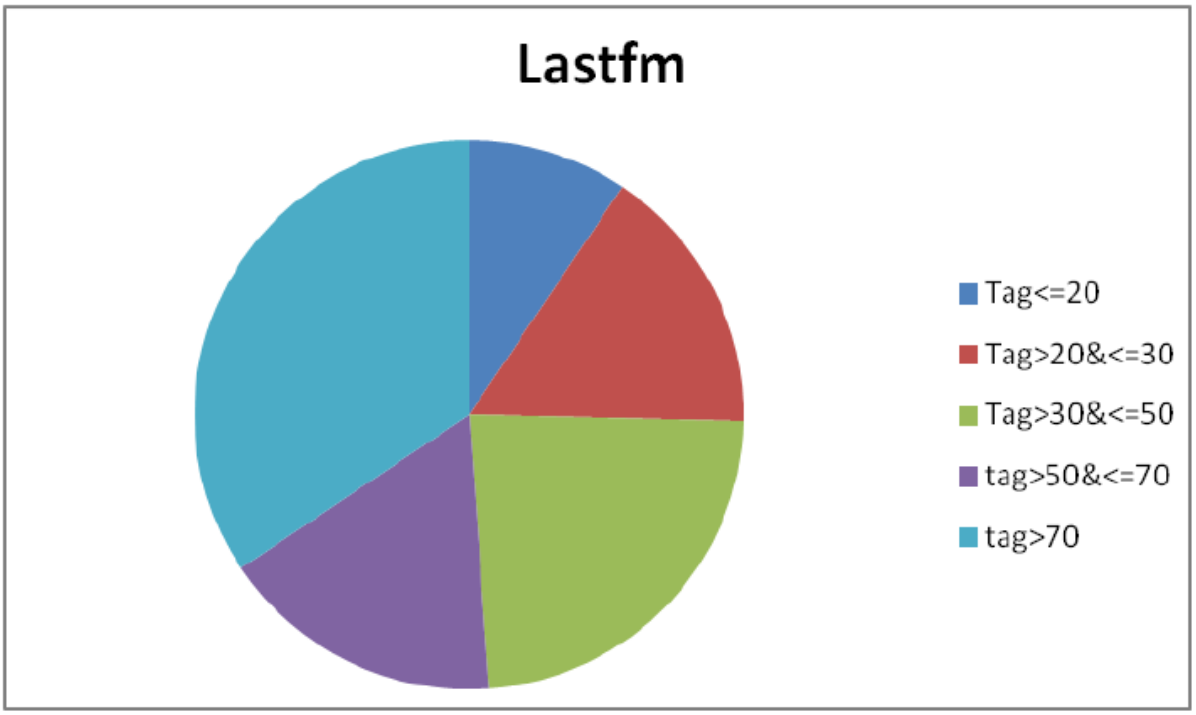

Figure 4. 10 - Distribution du nombre de tag attribué par utilisateur dans la base de données Lastfm.

Le tableau 4.1 présente une description complète des attributs de notre jeu de données.

Dataset	Last.fm
nombre d'utilisateurs	1892
Artistes	17632
Utilisateur-artiste	92834
Tags	11946
Utilisateurs-tags-artistes	186479
Utilisateur-utilisateur	92834

Tableau 4. 1 - Statistiques de données de hetrec-lastfm-2k dataset

Selon le tableau 4.1, nous expliquerons chaque caractéristique couverte dans cet ensemble de données.

- Artistes: Ce fichier contient des informations sur les artistes de musique écoutés et étiquetés par les utilisateurs. Dans ce fichier, une liste d'éléments se présente sous la forme:

  id_ artiste | *name_ artiste* | *url* | *photoURL*

  Exemple:

  707 Metallica http://www.last.fm/music/Metallica http://userserve-ak.last.fm/serve/252/7560709.jpg.

  59 New Order http://www.last.fm/music/New+Order http://userserve-ak.last.fm/serve/252/6650979.jpg.
- Tags : Ce fichier contient l'ensemble des tags disponibles dans l'ensemble de données. Dans ce fichier, une liste de tags se présente sous la forme de paires :

  tag_id | *tag_valeur*

  Exemple:

1	métal
4	black métal
62	trip hop

Tableau 4. 2 - Exemple de fichier de tag

- Utilisateur-artiste : Ce fichier contient les artistes écoutés par chaque utilisateur. Chaque ligne est un triple de :

  utilisateur-id | *artiste_id* | *poids*

Exemple :

78	2171	21
324	1325	311
2	81	1948

Tableau 4. 3 - Exemple de fichier d'utilisateur-artiste

- Utilisateurs-tags-artistes : Ce fichier contient les informations des utilisateurs et de leurs éléments sélectionnés et des tags associées. Chaque ligne se présente sous la forme :

  utilisateur-id | *artiste_id* | *tag_id* | *jour* | *mois* | *année*

  Exemple :

2	52	13	1	4	2009
299	1243	18	1	4	2008
304	72	74	1	7	2010

Tableau 4. 4 - Exemple de fichier d'annotation (Utilisateurs-tags-artistes)

- Utilisateur-utilisateur : Ces fichiers contiennent les relations amicales entre les utilisateurs de la base de données. Chaque ligne est une paire :

  utilisateur-id | *utilisateur-id*

  Ce qui signifie que ces deux utilisateurs sont des amis.

  Exemple :

11	816
174	320
2088	1688

Tableau 4. 5 - Exemple de fichier d'utilisateur-utilisateur (ami)

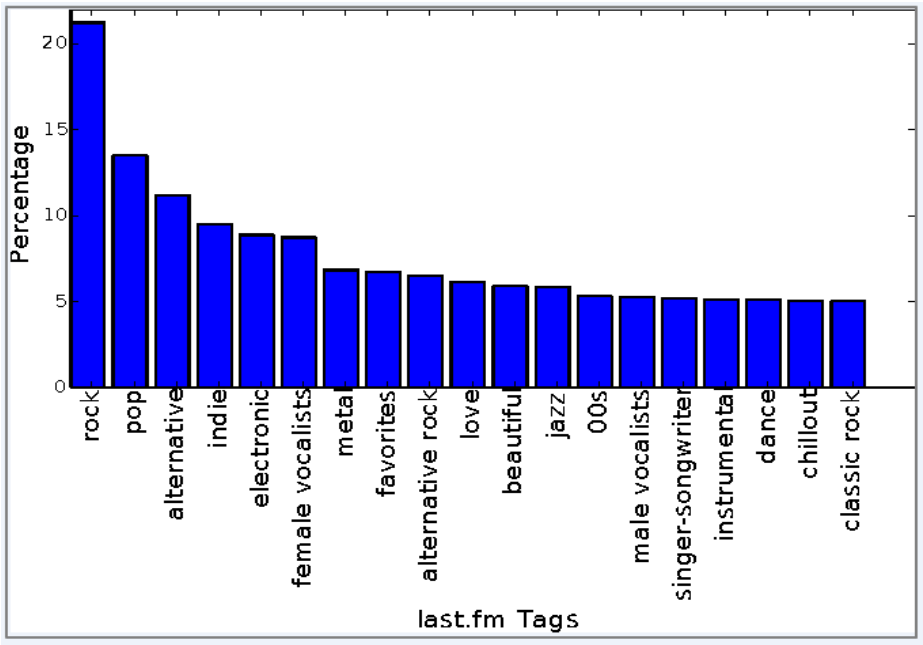

Figure 4. 11 - Distribution de tags pour les pistes

b) Division du jeu de données

La séparation des données en ensembles d'apprentissage et de test est une partie importante de l'évaluation des modèles d'exploration de données. En règle générale, lorsque vous séparez un jeu de données en un jeu d'apprentissage et un jeu de tests, la plupart des données sont utilisées pour l'apprentissage et une plus petite partie des données est utilisée pour les tests. Analyses Services échantillonne de manière aléatoire les données pour s'assurer que les tests et les ensembles d'apprentissage sont similaires. En utilisant des données similaires pour l'apprentissage et les tests, vous pouvez minimiser les effets des écarts de données et mieux comprendre les caractéristiques du modèle.

Une fois qu'un modèle a été traité à l'aide du jeu d'apprentissage, vous le testez en faisant des prédictions sur le jeu de test. Comme les données du jeu de test contiennent déjà des valeurs connues pour l'attribut que vous souhaitez prédire, il est facile de déterminer si les suppositions du modèle sont correctes.

À partir de l'ensemble de données initiales, nous avons divisé notre ensemble de données en deux parties: 80% de l'ensemble de données pour l'apprentissage et 20% de l'ensemble de données pour les tests. Dans la méthode de test 20-80, 80% de l'ensemble de données est sélectionné de manière aléatoire en tant qu'ensemble d'apprentissage et les 20% restants de l'ensemble de données sont sélectionnés en tant qu'ensemble de test.

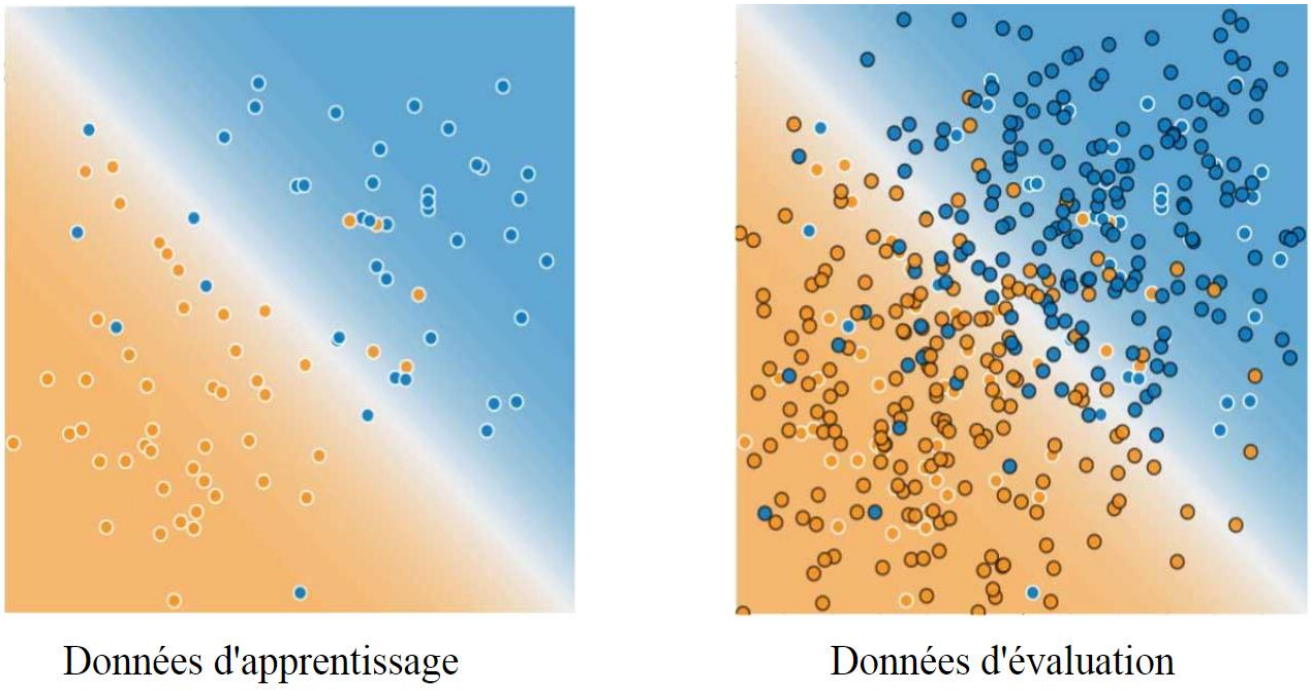

Figure 4. 12 - Division d'un ensemble de données en un ensemble d'apprentissage et un ensemble d'évaluation.

4.6. Métrique d'évaluation

1) MAE

Les mesures statistiques de précision consistent à évaluer la différence existante entre les notes prédites et les notes réellement attribuées par les utilisateurs. La MAE (*Mean Absolute Error*) est la mesure de précision la plus célèbre pour l'évaluation des systèmes de recommandation.

Cependant, la mesure MAE peut ne pas être appropriée pour l'évaluation de systèmes de recommandation qui génèrent des listes ordonnées de recommandation (listes TopN).

$$MAE = \frac{\sum_{(u,i)\in TestSet} |r_{u,i} - pr_{u,i}|}{|TestSet|} \qquad (4.10)$$

Où $pr_{u,i}$ représente la valeur d'évaluation prédite de l'élément i, et |TestSet| est le nombre de toutes les paires (u, i) dans l'ensemble de tests.

2) HMAE

La HMAE (High MAE) peut être utilisée pour évaluer la capacité d'un système de recommandation à proposer des items pertinents aux utilisateurs actifs. La différence entre HMAE et MAE, est que HMAE qui prend en considération uniquement les prédictions élevées.

M désigne le nombre d'items prédits avec des valeurs élevées. Plus la valeur de HMAE est faible, plus le système de recommandation est performant.

$$MAE = \frac{\sum_{(u,i)\in m} |r_{u,i} - pr_{u,i}|}{|m|} \qquad (4.11)$$

3) RMSE

RMSE est donné par **[109]** :

$$RMSE = \sqrt{\frac{\sum_{(u,i)\in TestSet}(r_{u,i} - pr_{u,i})^2}{|TestSet|}} \qquad (4.12)$$

4.7. Mesures d'évaluation utilisée

Afin de mesurer les taux d'erreur dans ce système de recommandation basé sur des tags, nous introduisons deux types de métriques d'évaluation qui sont les métriques les plus utilisées pour la précision de la prédiction. Nous utilisons les mesures de précision / rappel pour mesurer la performance des recommandations d'items. Ces deux métriques classiques sont également utilisées pour mesurer la qualité des tâches de récupération d'informations en général **[110]**.

1) Précision :

La précision est la proportion des items pertinents dans la liste des items retournés par le système, ce qui correspond au pourcentage ou au nombre d'items suggérés et s'avérant véritablement pertinent pour l'utilisateur actif. Si l'on considère par exemple, une liste des N. meilleures recommandations, la précision correspondra à la proportion d'items véritablement consultés et appréciés par l'utilisateur courant.

$$\text{précision} = \frac{|\{Recommandatio\ n\ pertinentes\}| \cap |\{Recommandatio\ n\ émises\}|}{|\{Recommandatio\ n\ émises\}|} \qquad (4.13)$$

Pour calculer la précision, nous avons utilisé une méthode qui consiste à considérer les TOP (K) films (les K meilleurs films). Cette méthode consiste à faire, pour un utilisateur donné, l'intersection entre les K premiers films notés effectivement par cet utilisateur et les K premiers films prédis par le prototype, et à diviser le résultat de cette intersection par le nombre K. La moyenne de ce calcul pour tous les utilisateurs nous délivre la précision pour la valeur de K. La

variation du nombre K nous permet d'obtenir d'autres valeurs de précision qui vont nous permettre de tracer la courbe de précision. La formule utilisée est la suivante :

$$\text{précision} = \frac{TopK\ prédits\ \cap Topk\ réels}{K} \qquad (4.14)$$

2) Rappel (Recall) :

Le rappel est la proportion des items pertinents par rapport à ceux qui sont pertinents au sein du corpus. Il compte le nombre d'items dans la liste de recommandations n-top ayant déjà été notés par l'utilisateur courant.

$$\text{Recall} = \frac{|\{Recommandatio\ n\ pertinentes\}| \cap\ |\{Recommandatio\ n\ émises\}|}{|\{Recommandatio\ n\ pertinentes\}|} \qquad (4.15)$$

Pour le calcul de cette métrique, nous procédons de la même façon avec laquelle nous avons calculé la précision. La formule du rappel est donnée par :

$$\text{Recall} = \frac{TopK\ prédits\ \cap Topk\ réels}{Topk\ réels} \qquad (4.16)$$

4.8. Outils de développements utilisés

Cette partie est consacrée à la présentation des différents outils et langages utilisés en justifiant nos choix techniques adoptés :

Les outils que nous avons utilisés pour le développement sont :

- Python version 2.7 comme des environnements de développement.
- Jeu de données : Lastfm-2k

4.8.1. Python

Python est un langage informatique de haut niveau, portable, dynamique, extensible, gratuit, structuré et open source conçu pour être orienté objet. Il est multi-paradigme et multi-usage. Développé à l'origine par *Guido Van Rossum* en 1993, Il est actuellement le langage le plus utilisé au monde. Comme il s'agit d'un langage de haut niveau, il est donc plus facile. Mais reste un langage très puissant.

4.8.2. Caractéristiques du langage

- Python convient aussi bien à des scripts d'une dizaine de lignes qu'à des projets complexes de plusieurs dizaines de milliers de lignes.
- Python est (optionnellement) multi-threadé.
- Python est portable, non seulement sur les différentes variantes d'UNIX, mais aussi sur les OS propriétaires : MacOS, BeOS, NeXTStep, MS-DOS et les différentes variantes de Window.
- Python est gratuit, mais on peut l'utiliser sans restriction dans des projets commerciaux
- Python gère ses ressources (mémoire, descripteurs de fichiers...) sans intervention du programmeur par un mécanisme de comptage de références (proche, mais différent d'un garbage collector).
- Python est un langage orienté-objet qui supporte l'héritage multiple et la surcharge des opérateurs.
- Python est dynamique (l'interpréteur peut évaluer des chaines de caractères représentant des expressions ou des instructions Python) orthogonal (un petit nombre de concepts suffit à engendrer des constructions très riches), réflectif (il supporte la méta programmation par exemple la capacité pour un objet de se rajouter ou de s'enlever des attributs ou des méthodes ou même de changer de classe en cours d'exécution) et introspectif (un grand nombre d'outils de développement comme le débugger ou le profiler sont implantés en Python lui-même)

4.9. Présentation de l'application

Après avoir justifié nos choix d'outils de développement, nous allons présenter quelques prises d'écrans de notre application.

Read Data :

Dans un premier temps, on va charger l'ensemble de données de lastfm, ainsi que les informations sur les caractéristiques extraites que ce soit pour les utilisateurs ou les items.

```
names = ['userID', 'artistID', 'tagID', 'timestamp']
path = 'hetrec2011-lastfm-2k/'
user_artist_tags_data = pd.read_csv(path,sep='\t', names=names)
```

Figure 4. 13 - Code source read data in pandas.

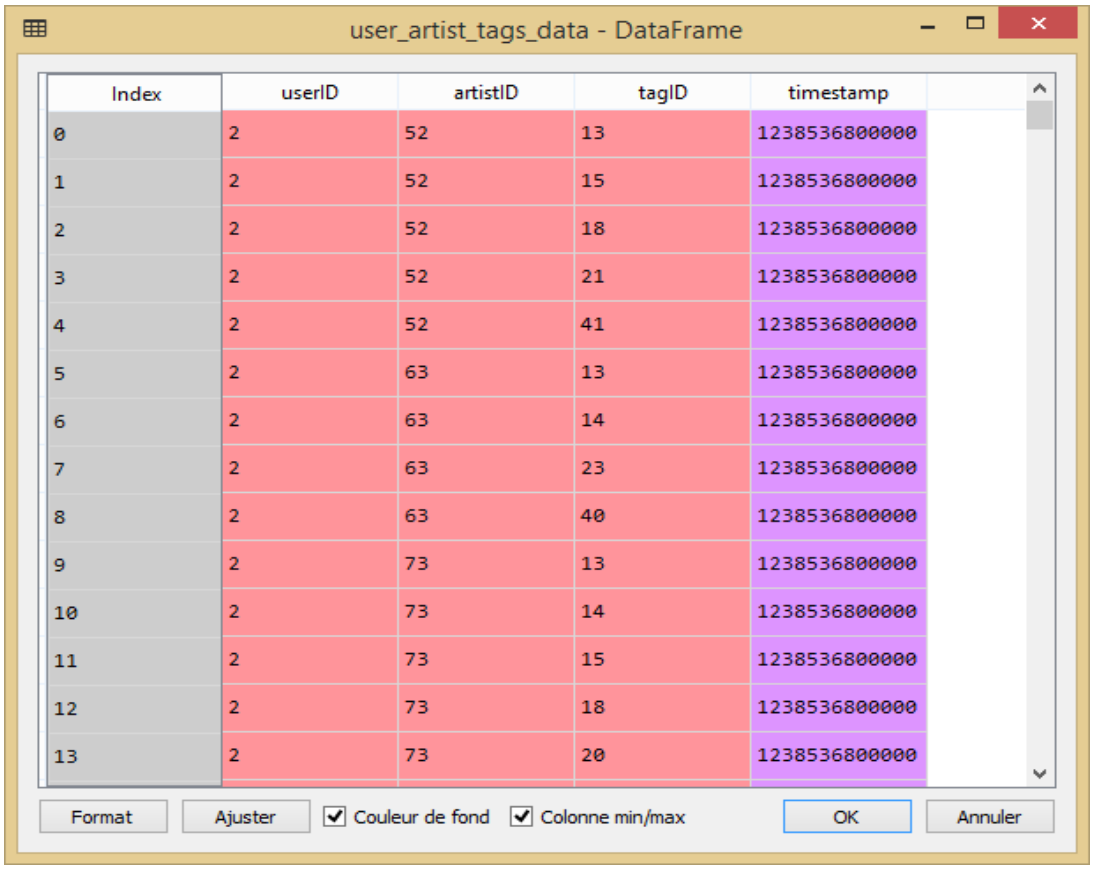

Figure 4. 14 - Fenêtre l'ensemble de données.

Division data :

```
from sklearn.model_selection import train_test_split
user_artist_tags_data, testSet = train_test_split(user_artist_tags_data1, test_size=0.2)
```

Figure 4. 15 - Code source Division data in pandas.

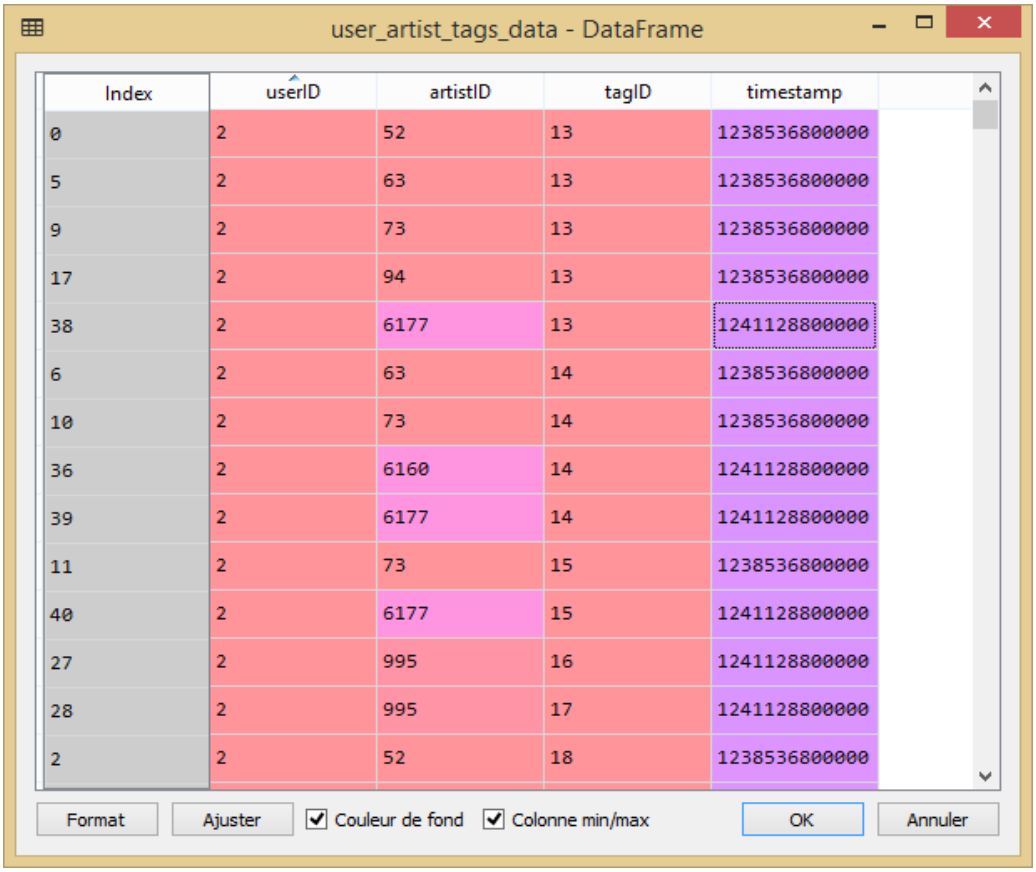

Figure 4. 16 - Fenêtre Training Data.

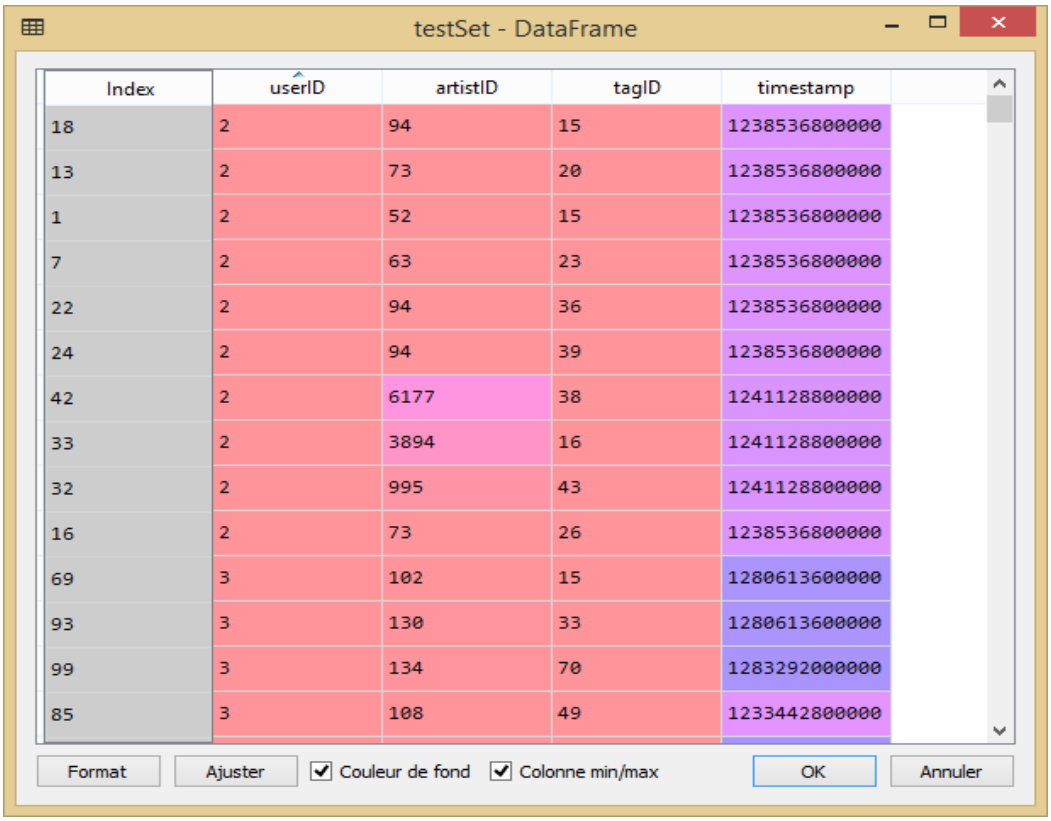

Figure 4. 17 - Fenêtre Testing Training Data.

Fenêtre Matrices : Une fois que l'utilisateur charge les fichiers, il pourra accéder à l'interface matrices pour observer les différentes matrices selon chaque ensemble.

```
1  uat_Sim_matrix = Sim_N.pivot_table(index='userID_i',columns='userID_j',values='Simu')
2  uat_Sim_matrix.as_matrix()
```

Figure 4. 18 - Code source convertie data en une matrice.

Figure 4. 19 - Fenêtre Matrice.

4.10. EXPÉRIENCES ET ANALYSE

Afin de mesurer les performances de notre système de recommandation proposé, certaines expériences ont été menées avec différents paramètres. Dans toutes les expériences, le jeu de données a été divisé en 20% de test et 80% de formation. L'objectif principal de la création du jeu de données d'apprentissage est de calculer la similarité des utilisateurs ainsi que les scores de prédiction. Après avoir calculé différents types de similarités, nous appliquons l'algorithme KNN pour sélectionner les items les plus proches et les utilisateurs les plus

similaires, respectivement pour l'item cible et l'utilisateur cible. Pour chaque utilisateur du jeu de données d'apprentissage, un score de prédiction sera calculé pour chaque item étiqueté par cet utilisateur (c'est-à-dire, les items ayant la note implicite 1). Les N premiers items seront recommandés à l'utilisateur. Ensuite, en utilisant le jeu de tests, la précision de l'algorithme peut être mesurée en comparant les items recommandés aux items que les utilisateurs ont réellement sélectionnés dans le jeu de tests. Ainsi, il est évident que moins la différence entre les jeux de données recommandés et le jeu de données déjà marqué (jeu de données test), plus l'algorithme proposé est précis. La précision et le rappel servent à évaluer l'exactitude des recommandations.

Nous pouvons également évaluer l'efficacité de l'approche proposée en comparant la précision et le rappel des principaux items recommandés de l'approche proposée à la performance des approches de filtrage collaboratif standard (FC) qui utilisent uniquement les informations sur l'item. Lorsque certains des poids de mesure de similarité sont mis à zéro. Par exemple, si α reçoit 1 alors que β vaut 0, cette approche devient un CF traditionnel basé sur l'utilisateur.

4.10.1. Trouver α et β

Pour l'algorithme basé utilisateur :

Pour obtenir les meilleures performances de notre solution proposée, un traitement préliminaire en termes d'initialisation de certaines variables est nécessaire. Selon l'équation (4.3), une échelle appropriée pour α et β est 0 -1.

Dans le sens de la découverte de la valeur la plus appropriée de α et β, nous modifions ces valeurs par incréments de 0,1 pour trouver les meilleures combinaisons de α et β, soit la combinaison présentant la valeur de précision la plus élevée. La valeur de précision la plus élevée indique que la plupart des éléments pertinents sont renvoyés à l'utilisateur, ce qui montre l'efficacité de l'algorithme proposée. Dans le tableau 4.6, nous avons examiné les valeurs de précision basées sur l'équation (4.14) tout en renvoyant les 5 recommandations principales. Le tableau 4.6 montre les valeurs de précision pour différentes combinaisons de α et β sur l'algorithme basée sur l'utilisateur, tandis que notre algorithme renvoie 5 items à un utilisateur donné.

β / α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	0.12258	0.12258	0.12220	0.12334	0.12315	0.12201	0.12163	0.12049	0.12068
0.2	0.12808	0.12713	0.12562	0.12600	0.12619	0.12657	0.12657	0.12751	0.12732
0.3	0.12865	0.12808	0.12789	0.12751	0.12638	0.12694	0.12657	0.12846	0.12846
0.4	0.12960	0.12960	0.12922	0.12979	0.12903	0.12770	0.12657	0.12789	0.12846
0.5	0.13055	0.13074	0.13017	0.13036	0.12998	0.13017	0.12884	0.12846	0.12827
0.6	0.13112	0.13074	0.13036	0.13036	0.13055	0.13074	0.12998	0.12979	0.12979
0.7	0.13112	0.13131	0.13150	0.13131	0.13074	0.13017	0.12979	0.13017	0.12960
0.8	0.13074	0.13074	0.13112	0.13112	0.13036	0.13017	0.12998	0.12998	0.12960
0.9	0.13093	0.13074	0.13055	0.13036	0.13036	0.13036	0.12998	0.12979	0.12979

Tableau 4. 6 - Valeur de précision pour les 5 top items.

Le tableau 4.6 montre que la performance de notre algorithme atteint son maximum lorsque α est égal à 0,7 et que β est égal à 0,3. Nous avons un tableau similaire pour rappel (Tableau 4.7) avec des résultats similaires pour α et β. Donc, ce seront les valeurs finales que nous utiliserons pour l'expérience ultérieure.

β / α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	0.18429	0.18328	0.18289	0.18368	0.18472	0.18285	0.18342	0.18189	0.18235
0.2	0.19062	0.18874	0.18757	0.18604	0.18582	0.18631	0.18657	0.18687	0.18692
0.3	0.19067	0.18978	0.18935	0.18921	0.18731	0.18823	0.18686	0.18792	0.18829
0.4	0.19149	0.19133	0.19117	0.19144	0.19015	0.18869	0.18647	0.18868	0.18878
0.5	0.19233	0.19256	0.19158	0.19142	0.19102	0.19163	0.18956	0.18906	0.18878
0.6	0.19252	0.19219	0.19192	0.19160	0.19217	0.19199	0.19069	0.19054	0.19044
0.7	0.19275	0.19279	0.19324	0.19300	0.19204	0.19127	0.19073	0.19112	0.19033
0.8	0.19249	0.19249	0.19281	0.19281	0.19143	0.19124	0.19102	0.19102	0.19049
0.9	0.19236	0.19189	0.19154	0.19135	0.19135	0.19135	0.19116	0.19092	0.19063

Tableau 4. 7 - Valeur de rappel pour les 5 top items.

Pour l'algorithme basé item :

Le tableau 4.8 montre que la performance de notre algorithme basé sur l'item qui atteint son maximum lorsque α est égal à 0,1 et que β est égal à 0,9. Donc, ce seront les valeurs finales que nous utiliserons pour l'expérience ultérieure

β / α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	0.0505	0.0750	0.0985	0.1192	0.1293	0.1361	0.1416	0.1444	0.1457
0.2	0.0554	0.0783	0.1006	0.1201	0.1289	0.1372	0.1392	0.1424	0.1448
0.3	0.0613	0.0821	0.1020	0.1197	0.1302	0.1343	0.1379	0.1392	0.1407
0.4	0.0681	0.0877	0.1033	0.1175	0.1293	0.1324	0.1363	0.1370	0.1390
0.5	0.0738	0.0902	0.1028	0.1158	0.1249	0.1298	0.1333	0.1350	0.1357
0.6	0.0799	0.0926	0.1026	0.1133	0.1230	0.1263	0.1287	0.1322	0.1337
0.7	0.0843	0.0932	0.1018	0.1101	0.1177	0.1204	0.1236	0.1289	0.1295
0.8	0.0864	0.0923	0.0972	0.1042	0.1087	0.1123	0.1153	0.1184	0.1208
0.9	0.0873	0.0891	0.0926	0.0967	0.1004	0.1020	0.1024	0.1061	0.1064

Tableau 4. 8 - Valeur de précision pour les 5 top items

Nous avons un tableau similaire pour rappel (Tableau 4.9) avec des résultats similaires pour α et β. Donc, ce seront les valeurs finales que nous utiliserons pour l'expérience ultérieure.

β / α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	0.0592	0.0894	0.1270	0.1612	0.1822	0.1937	0.2038	0.2068	0.2080
0.2	0.0635	0.0938	0.1295	0.1614	0.1808	0.1937	0.1983	0.2049	0.2069
0.3	0.0692	0.0981	0.1298	0.1607	0.1793	0.1896	0.1956	0.1985	0.2013
0.4	0.0779	0.1038	0.1309	0.1591	0.1776	0.1855	0.1933	0.1951	0.1999
0.5	0.0826	0.1069	0.1305	0.1548	0.1717	0.1811	0.1870	0.1920	0.1933
0.6	0.0920	0.1103	0.1303	0.1506	0.1668	0.1725	0.1772	0.1858	0.1884
0.7	0.0975	0.1104	0.1264	0.1435	0.1564	0.1640	0.1702	0.1756	0.1795
0.8	0.1004	0.1091	0.1170	0.1314	0.1393	0.1475	0.1540	0.1621	0.1659
0.9	0.1037	0.1062	0.1112	0.1175	0.1229	0.1255	0.1269	0.1354	0.1360

Tableau 4. 9 - Valeur de rappel dans le top 5

4.10.2. Résultats et analyse

Pour évaluer l'efficacité de la solution proposée, nous avons comparé la précision et le rappel des items recommandés du Top1et Top2 et Top5 et Top10.

Les résultats des valeurs de précision de notre algorithme basé sur l'utilisateur sont présentés dans le tableau 4.10, tandis que les résultats des valeurs de rappel sont énumérés dans le tableau 4.11.

Précision	**Top 1**	**Top 2**	**Top 5**	**Top 10**
Sim_Tcom	0.0843	0.1062	0.1041	0.0877
Sim_Tcom_Frie	**0.0923**	0.1133	0.1091	0.085
Sim_Tcom_Upsim	0.0891	0.1266	0.1278	0.1017
Sim_Global	0.0872	**0.1290**	**0.1315**	**0.1017**

Tableau 4. 10 -La précision sur l'algorithme basé sur l'utilisateur

Rappel	**Top 1**	**Top 2**	**Top 5**	**Top 10**
Sim_Tcom	0.0444	0.0868	0.1557	0.2135
Sim_Tcom_Frie	**0.0490**	0.0916	0.1618	0.2166
Sim_Tcom_Upsim	0.0453	0.1029	0.1894	0.2541
Sim_Global	0.0450	**0.1052**	**0.1932**	**0.2549**

Tableau 4. 11 - Le Rappel sur l'algorithme basé sur l'utilisateur

Le tableau 4.10 montre le résultat des valeurs d'évaluation sur la précision de notre algorithme basé sur l'utilisateur. Tout d'abord, cette valeur a été calculée lorsque la similarité était uniquement basée sur l'activité de marquage (***Sim_Tcom***). Deuxièmement, la valeur de précision a été calculée lorsque la similarité était basée sur la combinaison de l'activité de marquage avec les informations d'amitié (***Sim_Tcom_Frie***). Troisièmement, la précision était calculée lorsque la similarité était la combinaison de l'activité de marquage avec les informations des items (***Sim_Tcom_Upsim***). Enfin, la dernière valeur de précision a été calculée lorsque la similarité était la combinaison de ces trois types d'informations (***Sim_Global***). Nous pouvons voir des résultats similaires dans la valeur de rappel, comme indiqué dans le tableau 4.11.

La combinaison des trois similarités pourrait améliorer la précision de la recommandation en termes de valeur de précision. Parfois, les informations d'amitié donnent de meilleurs résultats et parfois les informations l'activité de marquage sont plus performantes. Globalement, lorsque nous recommandons 2 ou 5 items aux utilisateurs, la combinaison des trois

types d'informations nous donne les meilleurs résultats. Nous pouvons obtenir une conclusion similaire sur la valeur de rappel. La combinaison de des trois similarités peut également améliorer les performances sur les valeurs de rappel.

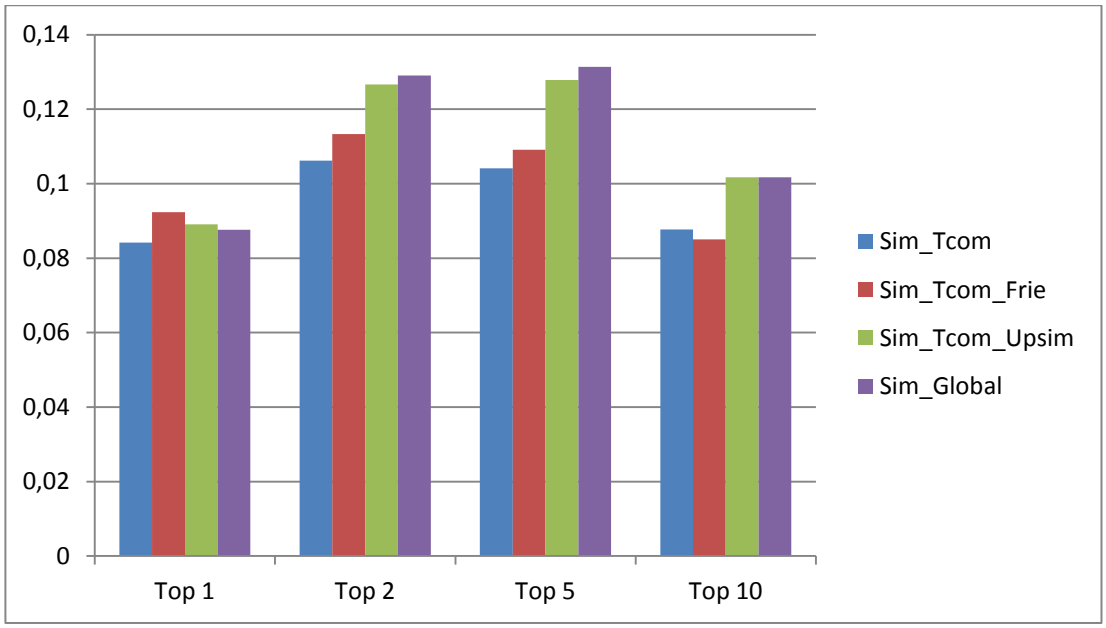

Figure 4. 20 - La précision sur l'algorithme basé sur l'utilisateur

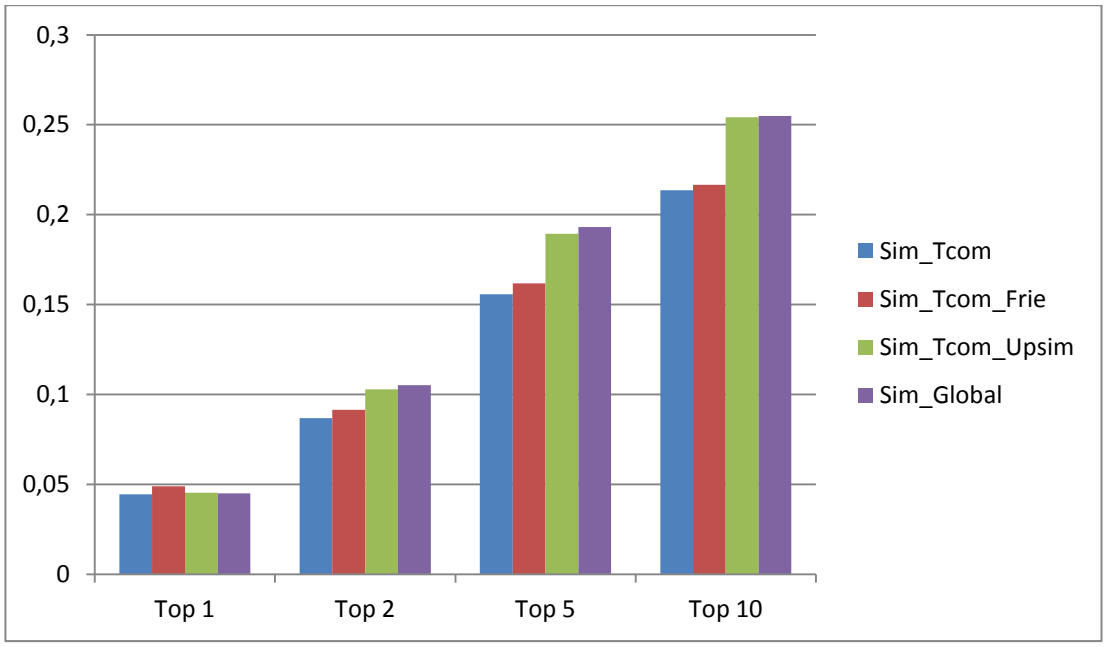

Figure 4. 21 - Le Rappel sur l'algorithme basé sur l'utilisateur

Les résultats des valeurs de précision de notre algorithme basé sur l'item sont présentés dans le tableau 4.12 tandis que les résultats des valeurs de rappel sont énumérés dans le tableau 4.13.

Précision	**Top 1**	**Top 2**	**Top 5**	**Top 10**
Sim_PUTSim	0.0280	0.0509	0.0854	0.0969
Sim_Tag_PUSim	0.0534	0.0893	0.1300	0.1356
Sim_Tag_PUTSim	0.0184	0.0322	0.0606	0.0700
Sim_Global	**0.0626**	**0.1045**	**0.1456**	**0.1515**

Tableau 4. 12 - La précision sur l'algorithme basé sur l'item

Rappel	**Top 1**	**Top 2**	**Top 5**	**Top 10**
Sim_PUTSim	0.0082	0.0315	0.0980	0.1723
Sim_Tag_PUSim	0.0234	0.0651	0.1823	0.2902
Sim_Tag_PUTSim	0.0055	0.0183	0.0653	0.1164
Sim_Global	**0.0290**	**0.0808**	**0.2080**	**0.3326**

Tableau 4. 13 - Le Rappel sur l'algorithme basé sur l'item

Le tableau 4.12 montre le résultat des valeurs d'évaluation sur la précision de notre algorithme basé sur l'item. Nous avons appliqué les mêmes étapes d'évaluation que celles appliquées dans l'algorithme basé sur l'utilisateur pour comparer les similitudes, et nous pouvons obtenir une conclusion similaire sur l'algorithme basé sur l'item. La combinaison de des trois similarités peut également améliorer les performances sur les valeurs de précision. Nous pouvons obtenir des résultats similaires dans la valeur de rappel, comme indiqué dans le tableau 4.13.

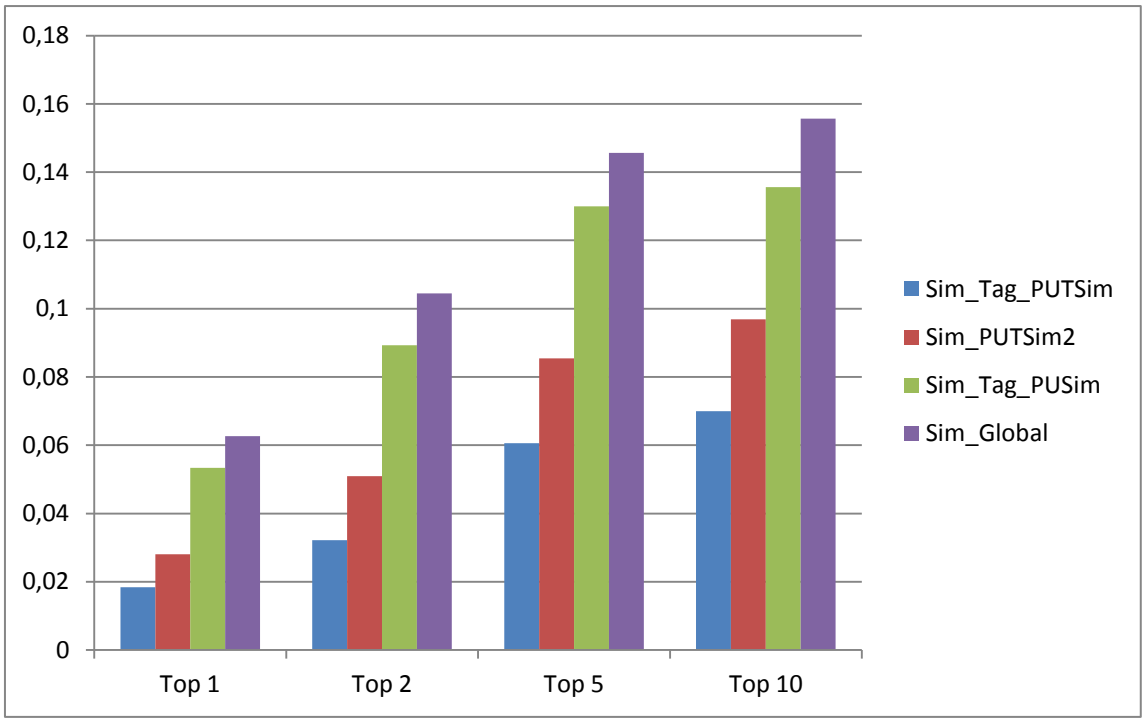

Figure 4. 22- La précision sur l'algorithme basé sur l'item

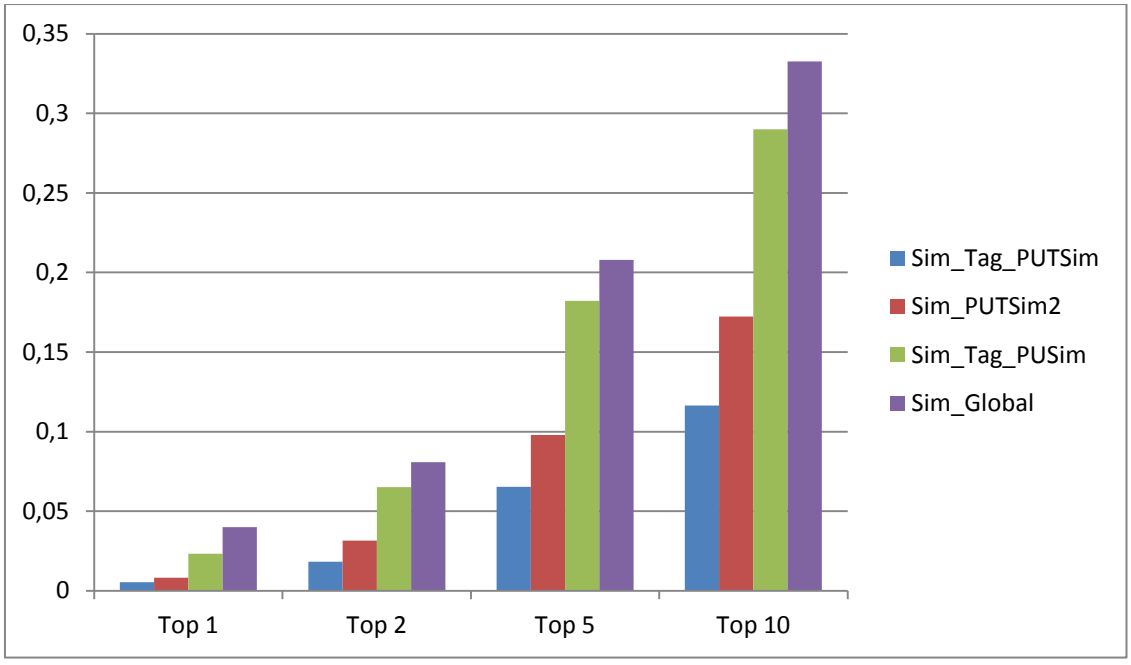

Figure 4. 23 -Le Rappel sur l'algorithme basée sur l'item

Les résultats des valeurs de précision de filtrage collaboratif traditionnel ainsi que notre solution proposée sur les deux algorithmes sont présentés dans le tableau 4.14.

	Précision	Top 1	Top 2	Top 5	Top 10
Algorithme basée sur l'utilisateur	**FC traditionnel**	0,0870	0,1190	0,1206	0,9317
	Solution proposée	0,0872	0,1290	0,1315	0,1017
Algorithme basée sur l'item	**FC traditionnel**	0,0377	0,0644	0,1095	0,1205
	Solution proposée	0,0626	0,1045	0,1456	0,1515

Tableau 4. 14 - Comparaison de la précision

En comparant les résultats de tableau 4.14, nous pouvons voir clairement que la solution proposée est bien meilleurs que le filtrage collaboratif traditionnel dans tous les aspects.

D'après les résultats du tableau 4.14, lorsque nous ajoutons des informations sociales telles que tags et l'amitié, la valeur de précision peut augmenter.

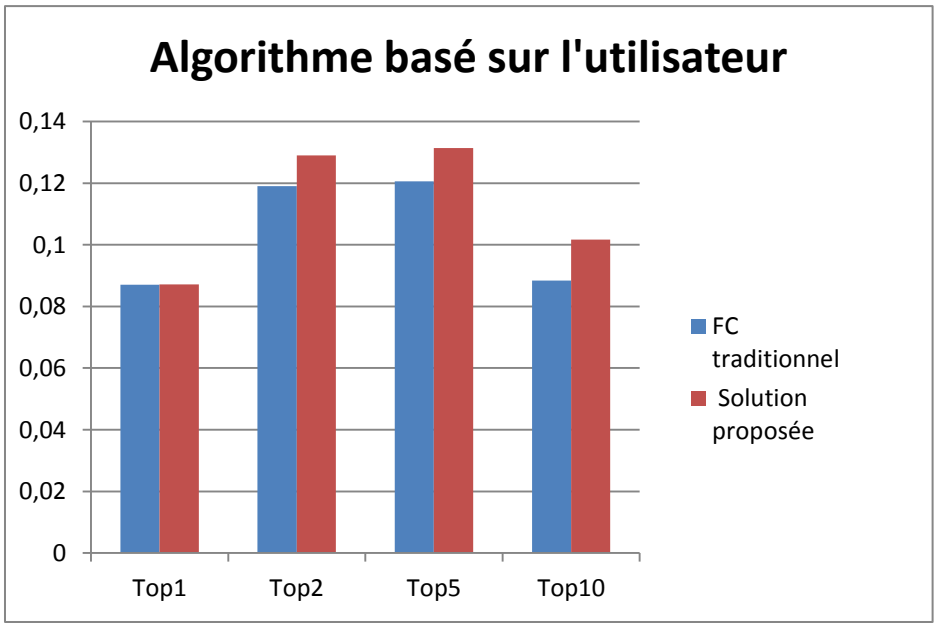

Figure 4. 24 - Comparaison de la précision pour l'algorithme basé sur l'utilisateur

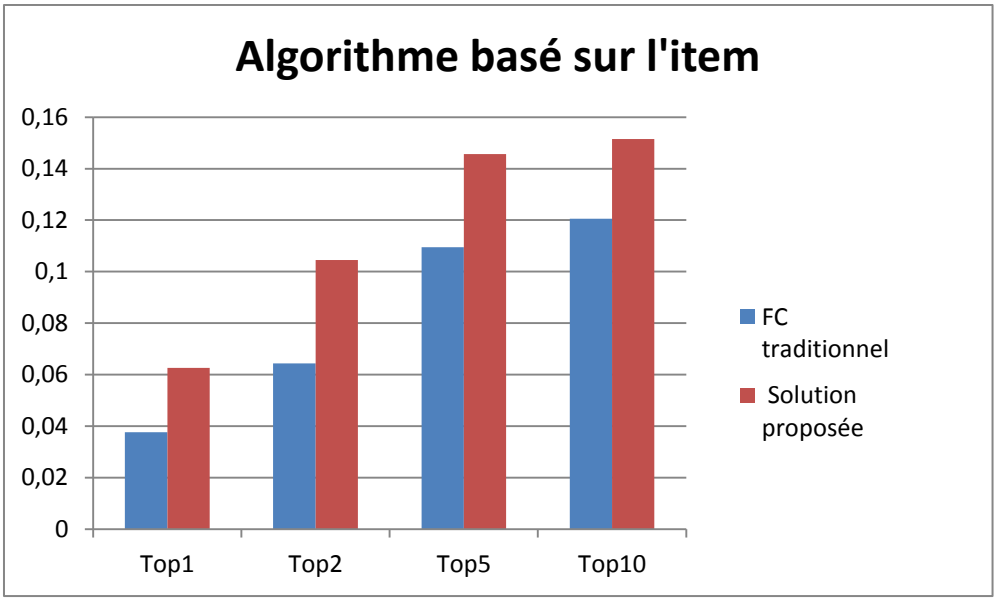

Figure 4. 25 - Comparaison de la précision pour l'algorithme basé sur l'item

Les résultats expérimentaux des figures 4.24 et 4.25 montrent que la précision de la solution proposée est supérieure à l'approche de filtrage collaboratif traditionnel pour les algorithmes basés sur l'utilisateur et les items.

Les résultats des valeurs de rappel de filtrage collaboratif traditionnel et notre solution sur les deux algorithmes sont présentés dans le tableau 4.15.

	Rappel	**Top 1**	**Top 2**	**Top 5**	**Top 10**
Algorithme basée sur l'utilisateur	**FC traditionnel**	0,0445	0,0966	0,1783	0,2322
	Solution proposée	0,045	0,1052	0,1932	0,2548
Algorithme basée sur l'item	**FC traditionnel**	0,0155	0,0443	0,1456	0,2620
	Solution proposée	0,029	0,0808	0,2080	0,3326

Tableau 4. 15 - Comparaison du rappel

Si on compare les résultats de tableau 4.15 on trouve que les résultats de la solution proposée sont meilleurs.

Les résultats des rappels présentés au tableau 4.15 montrent que la combinaison de l'activité de marquage de l'utilisateur avec les informations sociales peut également améliorer les performances sur les valeurs de rappel.

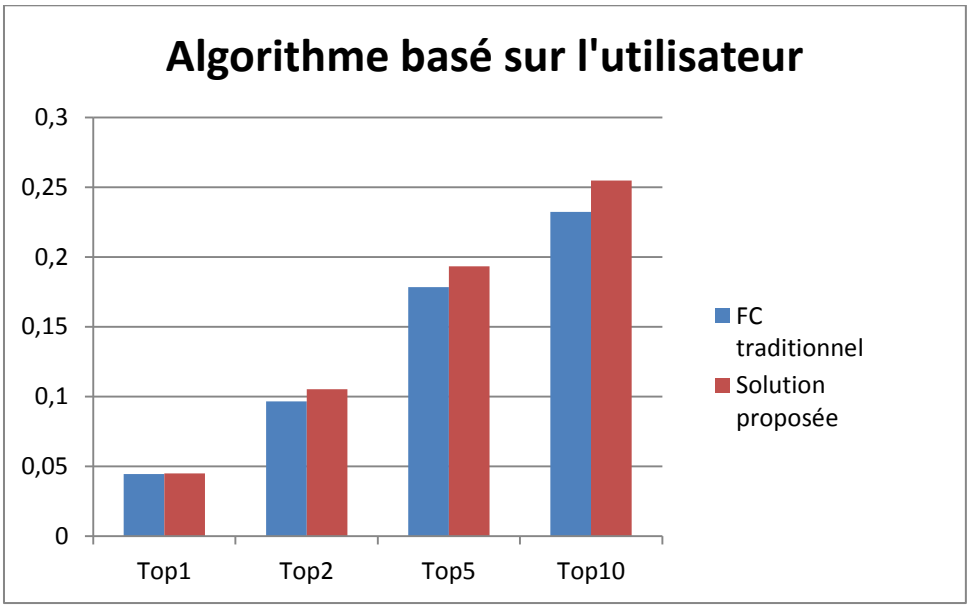

Figure 4. 26 - Comparaison du rappel pour le modèle basée sur l'utilisateur

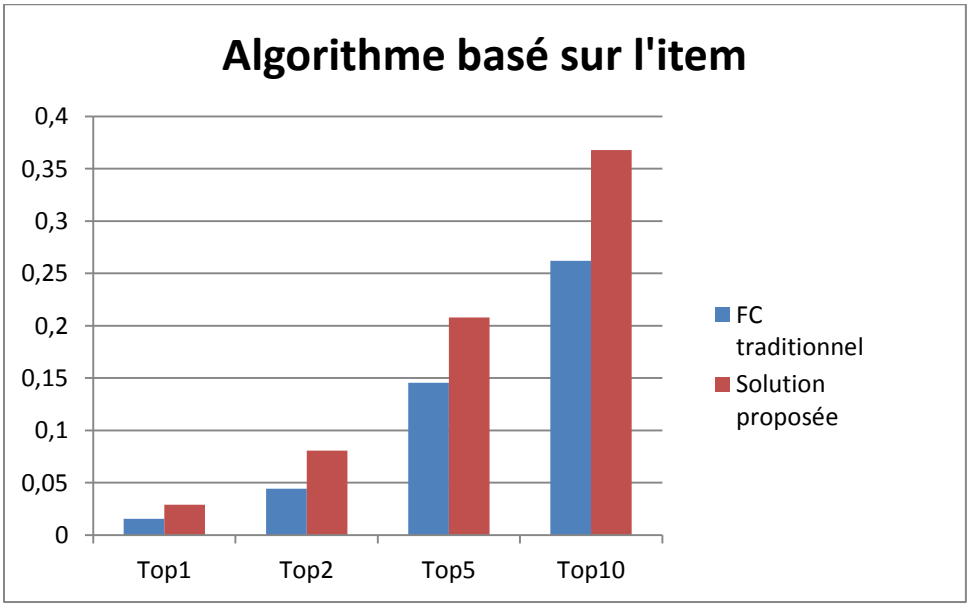

Figure 4. 27 - Comparaison du rappel pour l'algorithme basé sur l'item

Les résultats expérimentaux des figures 4.26 et 4.27 montrent que le rappel de la solution proposée est supérieur à l'approche de filtrage collaboratif traditionnel pour les deux algorithmes.

4.11. Conclusion

Dans ce chapitre, nous avons présenté notre proposition, les métriques d'évaluation, les outils de développement utilisés, ainsi que les résultats obtenus. Pour terminer, nous avons clôturé ce chapitre par une présentation de quelques prises d'écrans de notre application.

Conclusion Générale

Conclusion Générale

Les systèmes de recommandation (RS) visent à prédire les items ou les évaluations d'items qui intéressent l'utilisateur. Le filtrage collaboratif est une approche traditionnelle, largement utilisée pour recommander des items aux utilisateurs. Elle repose sur l'hypothèse que des personnes ayant les mêmes idées auront des goûts ou des comportements similaires. Cependant, les techniques de filtrage collaboratif couramment utilisées actuellement ne fonctionnent pas bien avec la relation tridimensionnelle distincte entre les utilisateurs, les tags et les items. En outre, la recommandation de filtrage collaboratif traditionnelle basée sur la similarité du comportement d'évaluation ne permet pas de traiter les informations de marquage collaboratif, et il présente aussi certaines limitations telles que le problème de démarrage à froid qui surgit lorsque des nouveaux items qui apparaissent et ne peuvent être point recommandé puisqu'ils n'ont reçu encore aucune note.

L'objectif principal de ce travail est comment utiliser les informations de tag et les informations de réseau social pour améliorer la qualité des systèmes de recommandation.

Nous avons proposé une méthode générique qui permet d'incorporer les tags à des algorithmes CF standard, en réduisant les corrélations tridimensionnelles à trois corrélations bidimensionnelles, puis en appliquant une méthode de fusion pour les associer à nouveau. Nous avons utilisé une métrique de similarité, basée sur les informations de marquage social, pour modéliser trois types de relations : Similarité de marquage, d'amitié et d'item. Les relations implicites sont conclues à partir de la méthode CF basée sur l'utilisateur en utilisant le comportement de l'utilisateur et l'activité de marquage en prenant en compte non seulement les items partagés par les utilisateurs, mais également les tags partagées par les utilisateurs. En d'autres termes, nos informations implicites sont déduites de tags partagées sur des items partagés. Des informations explicites sont recueillies à partir des relations sociales des utilisateurs, y compris de leur amitié. Les évaluations empiriques sur des algorithmes de FC avec un ensemble de données réelles démontrent que l'intégration des tags dans notre approche proposée fournit des résultats prometteurs et significatifs.

Bien que l'utilité de notre approche proposée ait été démontrée dans notre évaluation à l'aide de données réelles. Les méthodes particulières que nous avons adoptées pour formuler des recommandations et améliorer leur précision sont de nature heuristique. Comment améliorer la précision de la recommandation reste une question ouverte. Ainsi, notre travail ouvre plusieurs pistes pour des recherches futures. Dans un sens, il serait extrêmement intéressant d'étudier l'utilisation des informations sémantiques de ces tags. Ainsi, nous pouvons étendre notre approche à une nouvelle méthode basée sur la sémantique avec une approche hybride qui utilise

une combinaison de filtrage CF et de filtrage basé sur le contenu pour vérifier si elle pourrait améliorer davantage les performances. Cela signifie que nous devons analyser la signification sémantique et le contexte des tags sociales pour rechercher les utilisateurs similaires ou des items similaires **[111]**. Une autre direction que nous aimerions envisager consiste à élargir notre approche pour intégrer la dimension temporelle en tant que mesure permettant d'évaluer l'importance d'une paire de tags d'item. De plus, nous pouvons faire avancer l'utilisation des relations d'amitié en considérant la relation transitive (Amis les plus proches d'un ami) entre les utilisateurs. Par conséquent, un nouveau graphique peut être créé, lequel définit des relations plus larges entre les utilisateurs.

Bibliographie et Références

Bibliographie Et Références

[1] Samaidhu,(January 24,2015), recommendation engine , https://dataaspirant. com/2015/01/24/recommendation-engine-part-1/.

[2] Burke, R. (2002). Hybrid recommender systems : Survey and experiments. User Mo-deling and User-Adapted Interaction.

[3] Mathieu, 25.04.2012, in Dossiers, Mathieu Les algorithmes de re-commandation https://www.podcastscience.fm/dossiers/2012/04/25/les-algorithmes-de-recommandation/ consulté le 03/06/2019

[4] An Te NGUYEN, thèse Doctorat : «COCoFil2 : Un nouveau système de filtrage collaboratif basé sur le modèle des espaces de communautés», Université Joseph Fourier-Grenoblei, 23/11/2006.

[5] Catherine B ERRUT et Nathalie D ENOS, chapitre 8 «Filtrage collaboratif», p 242-268, E. GAUSSIER, M.H. STEFANINI, « Assistance intelligente à la RI », Hermes-Lavoisier, 2003.

[6] NEGRE Elsa, projet de recherche: «Les systèmes de recommandation», Université Paris-Dauphine, 2008.

[7] Imane BOUSSEBOUGH ép.BOUGHERRA, thèse Doctorat : «Les systèmes multi-agents dynamiquement adaptables », département informatique, Université Mentouri Constantine, 06/07/2011.

[8] M. Amokrane BELLOUI, thèse magister: «L'usage des concepts du web sémantique dans le filtrage d'information collaboratif», Ministère de l'enseignement Supérieur et de la recherche scientifique Institut National d'Informatique Alger, 2008.

[9] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J. (1994). Grouplens :An open architecture for collaborative filtering of netnews. In Proceedings of the1994 ACM Conference on Computer Supported Cooperative Work, CSCW '94, pages175–186, New York, NY, USA. ACM.

[10] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001). Item-based collaborativefiltering recommendation algorithms. In Proceedings of the 10th InternationalConference on World Wide Web, WWW '01, pages 285–295, New York, NY, USA.ACM.

[11] Berrut 03 C. Berrut, "Filtrage collaboratif", chapitre 8, p 255-283, E. GAUSSIER, M.H. STEFANINI, "Assistance intelligente à la recherche d'information", Hermes-Lavoisier, 2003.

[12] Maria Lizpbeth Gallardo Lopez, « Accés à l'information par un système de filtrage collaboratif contrôlé », thèse de doctorat, université de Joseph Fourier, 2005.

[13] Nicolas Lumineau, « Un tour d'horizon du filtrage collaboratif », Travail réalisé dans le cadre de l'AS Personnalisation de l'information, Laboratoire d'informatique de Paris 6, 2002.

[14] Samia BOULKRINAT, « Modélisation hybride du profil utilisateur pour un système de filtrage d'informations sur le web », thème du mémoire, Magistère Ingénierie des Systèmes d'Information et Document Electronique, INI, 2007.

[15] Laurent Candillier, « Apprentissage automatique de profils de lecteurs », mémoire de DEA, Laboratoire d'Informatique Fondamentale de Lille, Juin 2001.

[16] Nguyen, A. T. (2006). COCoFil2 : Un nouveau système de filtrage collaboratif basé sur le modèle des espaces de communautés. PhD thesis, université Joseph Fourier-Grenoble I.

[17] Arnautu, O. R. (2012). Mures : Un système de recomman-dation de musique. Master's thesis, La Faculté des arts et des sciences Université de Montréal.

[18] Wang, C. and Blei, D. M. (2011). Collaborative topic modeling for recommending scienti c articles. In Proceedings of the 17th ACM SIGKDD International Confe-rence on Knowledge Discovery and Data Mining, KDD '11, pages 448-456, New York, NY, USA. ACM.

[19] Wetzker, R., Zimmermann, C., and Bauckhage, C. (2008). Analyzing Social Bookmar-king Systems : A del.icio.us Cookbook. In Proceedings of the ECAI 2008 Mining Social Data Workshop, pages 26-30. IOS Press

[20] Huang, C.-L. and Lin, C.-W. (2010). Collaborative and content-based recommender system for social bookmarking website. World Academy of Science, Engineeringand Technology, 4(8) :606 – 611.

[21] Liu, F. and Lee, H. J. (2010). Use of social network information to enhance collaborative filtering performance. Expert Systems with Applications, 37(7) :4772 –4778.

[22] Zheng, N. and Li, Q. (2011). A recommender system based on tag and time informa-tion for social tagging systems. Expert Syst. Appl., 38(4) :4575-4587.

[23] Brusilovsky, P., Cassel, L. N., Delcambre, L. M., Fox, E. A., Furuta, R., Garcia, D. D., III, F. M. S., and Yudelson, M. (2010). Social navigation for educational digital libraries. Procedia Computer Science, 1(2) :2889 { 2897. Proceedings of the 1st Workshop on Recommender Systems for Technology Enhanced Learning (RecSysTEL 2010) Proceedings of the 1st Workshop on Recommender Systems for Technology Enhanced Learning (RecSysTEL 2010).

[24] Beldjoudi, S., Seridi, H., and Zucker, C. F. (2011). Improving tag-based resource recommendation with association rules on folksonomies. In In Proceedings of the 11th Interational Semantic Web Conference ISWC2011.

[25] Kim, H.-N., Roczniak, A., Levy, P., and Saddik, A. (2012). Social media ltering based on collaborative tagging in semantic space. Multimedia Tools Appl., 56(1) :63-89.

[26] Joly, A., Maret, P., and Daigremont, J. (2010). Contextual recommendation of social updates, a tag-based framework. In Proceedings of the 6th International Conference on Active Media Technology, AMT'10, page 436-447, Berlin, Heidelberg. Springer-Verlag.

[27] Manzat, A.-M., Grigoras, R., and Sedes, F. (2010). Towards a user-aware enrichment of multimedia metadata. In Workshop on Semantic Multimedia Database Technologies (SMDT 2010), pages 30{41, Saarbr•ucken, Germany. CEUR Workshop Proceedings.

[28] Mishne, G. (2006). Autotag : A collaborative approach to automated tag assignmentfor weblog posts. In Proceedings of the 15th International Conference on WorldWide Web, WWW '06, pages 953–954, New York, NY, USA. ACM.

[29] Sood, S. C. and Hammond, K. J. (2007). Tagassist : Automatic tag suggestion for blog posts. In In International Conference on Weblogs and Social.

[30] Musto, C., Narducci, F., de Gemmis, M., Lops, P., and Semeraro, G. (2009). Star : a social tag recommender system. In Eisterlehner, F., Hotho, A., and J•aschke, R., editors, ECML PKDD Discovery Challenge 2009 (DC09), volume 497, pages 215{227, Bled, Slovenia. CEUR Workshop Proceedings

[31] Le Deuff.O: Folksonomies, (2006) Les usagers indexent le web. BBF -Paris, t. 51, n° 4.

[32] Golder S., Bernardo A., Huberman (Aug, 2005).The Structure of Collaborative Tagging Systems. Journal of Information Science32 (2):198--208.

[33] Durieux,V (2010). Collaborative tagging et folksonomies, L'organisation du web par les internautes, Les Cahiers du numérique, 2010/1Vol. 6, p. 69-80.

[34] Panke S., Gaiser S., (2009). With My Head Up in the Clouds : Using Social Tagging to Organize Knowledge , Journal of Business and Technical Communication, vol. 23, n° 3, , p. 318-349.

[35] Golder S.A., Huberman B.A., (2006) « Usage patterns of collaborative tagging systems », Journal of Information Science, vol. 58, n° 8, p. 1175-1187.

[36] Marlow C., Naaman M., Boyd D., Davis M. (août 2006) HT06, Tagging Paper, Taxonomy, Flickr,Academic Article, ToRead , Proceedings of the seventeenth conference on Hypertext and Hypermedia, Odense, 22-25, New-York, ACM, p. 31-40

[37] Halpin H., Robu.V, Shepherd.H. (2007). The Complex Dynamics of Collaborative Tagging. In WWW : ACM Press.

[38] Gruber.T (2007) Ontology of folksonomy: A mash-up of apples and oranges. Int. Journal on Semantic Web and Information Systems.

[39] VanderWal, T. (2005) – Explaining and Showing Broad and Narrow Folksonomies. http://www.vanderwal.net/random/entrysel.php?blog=1635.

[40] Vander Wal, T. (2007). Folksonomy coinage and definition.

[41] Broudoux.E : Folksonomie et indexation collaborative, rôle des réseaux sociaux dans la fabrique de l'information. Collaborative Web Tagging Workshop at WWW 2006, Edinburgh, Scotland, May, 2006.

[42] http://www.Wikipedia.com

[43] Beldjoudi, S., Seridi, H., & Faron-Zucker, C. (2012, November). Let Tagging be More Interesting. In Advanced Information Systems for Enterprises (IWAISE), 2012 Second International Workshop on (pp. 2-8). IEEE.

[44] Weinberger, D (2007). Everything is miscellaneous: the power of the new digital disorder. New York: Times BOOK.

[45] Gueddari K. (2009). Folksonomie: la gestion collaboratives des signets.

[46] Mathes, A. (2004). Folksonomies-cooperative classification and communication through shared metadata.

[47] Mejías, U. A. (2007). Tag literacy. ideant:.

[48] Limpens, F., Gandon, F., & Buffa, M. (2009). Sémantique des folksonomies: structuration collaborative et assistée. In Ingénierie des Connaissances (pp. 37-48).

[49] Specia, L., & Motta, E. (2007). Integrating folksonomies with the semantic web. In The semantic web: research and applications (pp. 624-639). Springer Berlin Heidelberg.

[50] Cattuto, C., Benz, D., Hotho, A., & Stumme, G. (2008). Semantic grounding of tag relatedness in social bookmarking systems (pp. 615-631). Springer Berlin Heidelberg.

[51] Zacklad, M. (2007, May). Classification, thésaurus, ontologies, folksonomies: comparaisons du point de vue de la recherche ouverte d'information (ROI). InCAIS/ACSI 2007, 35e Congrès annuel de l'Association Canadienne des Sciences de l'Information. Partage de l'information dans un monde fragmenté: Franchir les frontières, sous la dir. de C. Arsenault et K. Dalkir. Montréal: CAIS/ACSI, 2007.

[52] Limpens, F., Gandon, F., & Buffa, M. (2008). Rapprocher les ontologies et les folksonomies pour la gestion des connaissances partagées: un état de l'art. In19es Journées Francophones d'Ingénierie des Connaissances (IC 2008) (pp. 123-134).

[53] Passant, A. (2007, March). Using ontologies to strengthen folksonomies and enrich information retrieval in weblogs. In International Conference on Weblogs and Social Media.

[54] Gruber, T. (2008).Collective knowledge systems: Where the social web meets the semantic web. Web semantics: science, services and agents on the World Wide Web, 6(1),4-13.

[55] Mika, P. (2005). Ontologies are us: A unified model of social networks and semantics. In The Semantic Web–ISWC 2005 (pp. 522-536). Springer Berlin Heidelberg.

[56] Bischoff, K., Firan, C. S., Nejdl, W., & Paiu, R. (2008, October). Can all tags be used for search?. In Proceedings of the 17th ACM conference on Information and knowledge management (pp. 193-202). ACM.

[57] Xu, S., Bao, S., Fei, B., Su, Z., & Yu, Y. (2008, July). Exploring folksonomy for personalized search. In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 155-162). ACM.

[58] Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B., & Su, Z. (2007, May). Optimizing web search using social annotations. In Proceedings of the 16th international conference on World Wide Web (pp. 501-510). ACM.

[59] Millen, D., Feinberg, J., Kerr, B., « Dogear: Social Bookmarking in the enterprise », Proceedings of the 24th international conference on Human Factors in compu ting systems (CHI 2006), Montreal, Canada, 22-27 avril 2006, p. 111-120.

[60] Kalamatianos, A., Zervas, P., Sampson, D.G., « ASK-LOST 2.0: A Web-Based Tool for Social Tagging of Digital Educational Resources », Proceedings of the Ninth IEEE International Conference on Advanced Learning Technologies (ICALT 2009), Riga, Latvia, 15-17 juillet 2009, p. 157-159.

[61] Conole, G., .Culver, J., « The design of Cloudworks: Applying social networking practice to foster the exchange of learning and teaching ideas and designs », Computers & Education, vol. 54, n° 3, 2010, p. 679-692.

[62] Huynh-Kim-Bang, B., Indexation de documents pédagogiques : fusionner les approches du Web Sémantique et du Web Participatif, Thèse de doctorat, Université Henri Poincaré - Nancy I, 2009, 274p.

[63] Robert J¨aschke, Leandro Marinho, Andreas Hotho, Lars Schmidt-Thieme, and Gerd Stumme.Tag recommendations in social bookmarking systems.AI Communications, pages 231–247,2008.16. Robe

[64] Andriy Shepitsen, Jonathan Gemmell, Bamshad Mobasher, and Robin Burke. Personalizedrecommendation in social tagging systems using hierarchical clustering. In RecSys ’08: Pro-ceedings of the 2008 ACM conference on Recommender systems, pages 259–266, New York,NY, USA, 2008. ACM.

[65] Tchuente D., Canut M.-F., Jessel N., Peninou A., Sèdes F. (2013). A community-based algorithm for deriving users’ profiles from egocentrics networks: experiment on facebook and DBLP. Social Network Analysis and Mining, vol. 3, no 3, p. 667–683. Consulté sur http://link.springer.com/article/10.1007/s13278-013-0113-0

[66] Ma Y., Zeng Y., Ren X., Zhong N. (2011). User interests modeling based on multi-source personal information fusion and semantic reasoning. In Proceedings of the 7th international conference on active media technology, p. 195–205. Berlin, Heidelberg, Springer-Verlag. Consulté sur http://dl.acm.org/citation.cfm?id=2033896.2033923

[67] Meo, P. D., Quattrone, G., and Ursino, D. (2010). A query expansion and user profile enrichment approach to improve the performance of recommender systems operating on a folksonomy. User Modeling and User-Adapted Interaction, 20(1) :41–86.

[68] Kim, H.-N., Alkhaldi, A., El Saddik, A., and Jo, G.-S. (2011). Collaborative user modeling with user-generated tags for social recommender systems. Expert Systems with Applications, 38(7) :8488–8496.

[69] White, R. W., Bailey, P., and Chen, L. (2009). Predicting user interests from contextual information. In Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’09, page 363–370, New York, NY, USA. ACM.

[70] Godoy, D. and Amandi, A. (2008). Hybrid content and tag-based profiles for recommendation in collaborative tagging systems. In Latin American Web Conference, 2008. LA-WEB ’08., pages 58–65.

[71] Cai, Y. and Li, Q. (2010). Personalized search by tag-based user profile and resource profile in collaborative tagging systems. In Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM ’10, pages 969–978, New York, NY, USA. ACM

[72] Song, Y., Zhang, L., and Giles, C. L. (2011). Automatic tag recommendation algorithms for social recommender systems. ACM Trans. Web, 5(1) :4 :1–4 :31.

[73] Astrain, J. J., Cordoba, A., Echarte, F., and Villadangos, J. (2010). An algorithm for the improvement of tag-based social interest discovery. In SEMAPRO ’10 : Proceedings of The Fourth International Conference on Advances in Semantic Processing, pages 49–54.

[74] Musia l, K. and Kazienko, P. (2013). Social networks on the internet. World Wide Web, 16(1) :31–72.

[75] Tchuente, D., Canut, M.-F., Jessel, N. B., P´eninou, A., and S`edes, F. (2012). Visualizing the relevance of social ties in user profile modeling. Web Intelligence and Agent Systems, 10(2) :261–274.

[76] Li, R., Wang, C., and Chang, K. C.-C. (2014b). User profiling in an ego network : Co-profiling attributes and relationships. In Proceedings of the 23rd International Conference on World Wide Web, WWW '14, pages 819–830, New York, NY, USA. ACM.

[77] C. Shiky. Ontology is overrated http://www.shirky.com/ontology_overrated.html, 2005.Retrieved on May 26, 2005.retrieved on May 26, 2007

[78] Marinho, L.B. and Schmidt-Thieme, L.: Collaborative tag recommendations: Data Analysis, Machine Learning and Applications. In the 31st Annual Conference of the Gesellschaft für Klassifikation. pp. 533-540.Springer, Berlin Heidelberg (2007)

[79] M. Braunhofer, M. Kaminskas, and F. Ricci, "Recommending music for places of interest in a mobile travel guide," in Proceedings of the fifth ACM conference on Recommender systems, pp. 253-256, 2011.

[80] Zhao, S., Du, N., Nauerz, A., Zhang, X., Yuan, Q., and Fu, R. (2008). Improved recommendation based on collaborative tagging behaviors. In Proceedings of the 13th International Conference on Intelligent User Interfaces, IUI '08, pages 413– 416, New York, NY, USA. ACM.

[81] Cabanac, G. (2011). Accuracy of inter-researcher similarity measures based on topical and social clues. Scientometrics, 87(3) :597–620.

[82] Guy, I., Zwerdling, N., Ronen, I., Carmel, D., and Uziel, E. (2010). Social media recommendation based on people and tags. In Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '10, pages 194–201, New York, NY, USA. ACM.

[83] Roth, M., Ben-David, A., Deutscher, D., Flysher, G., Horn, I., Leichtberg, A., Leiser, N., Matias, Y., and Merom, R. (2010). Suggesting friends using the implicit social graph. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10, pages 233–242, New York, NY, USA. ACM.

[84] Canut, C. M., Mezghani, M., On-at, S., P´eninou, A., and S`edes, F. (2015). A comparative study of two egocentric-based user profiling algorithms - experiment in delicious. In ICEIS 2015 - Proceedings of the 17th International Conference on Enterprise Information Systems, Volume 2, Barcelona, Spain, 27-30 April, 2015, pages 632–639.

[85] Gupta, M., Li, R., Yin, Z., and Han, J. (2010). Survey on social tagging techniques. SIGKDD Explor. Newsl., 12(1) :58–72.

[86] Nauerz, A., Pietschmann, S., and Pietzsch, R. (2008). Social recommendation and adaptation in web portals. In Nejdl, W., Kay, J., Pu, P., and Herder, E., editors, Proceedings of the Workshop on ”Adaptation for the Social Web” ASW. Springer.

[87] Michlmayr, E. (2007). Learning user profiles from tagging data and leveraging them for personal(ized) information access. In In Proceedings of the Workshop on Tagging and Metadata for Social Information Organization, 16th International World Wide Web Conference (WWW2007).

[88] Huang, C.-L., Chien, H.-Y., and Conyette, M. (2011). Folksonomy-based recommender systems with user-s recent preferences. 5(6) :127 – 131.

[89] Tso-Sutter, K. H. L., Marinho, L. B., and Schmidt-Thieme, L. (2008). Tag-aware recommender systems by fusion of collaborative filtering algorithms. In Proceedings of the 2008 ACM Symposium on Applied Computing, SAC ’08, pages 1995–1999, New York, NY, USA. ACM.

[90] Wang, J., Clements, M., Yang, J., de Vries, A. P., and Reinders, M. J. T. (2010). Personalization of tagging systems. Inf. Process. Manage., 46(1) :58–70.

[91] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions," Knowledge and Data Engineering, IEEE Transactions on, vol. 17, pp. 734-749, 2005.

[92] White, R. W., Bailey, P., and Chen, L. (2009). Predicting user interests from contextual information. In Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’09, page 363–370, New York, NY, USA. ACM.

[93] Vallet, D., Cantador, I., and Jose, J. M. (2010). Personalizing web search with folksonomy-based user and document profiles. In Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., R¨uger, S., and Rijsbergen, K. v., editors, Advances in Information Retrieval, number 5993 in Lecture Notes in Computer Science, pages 420–431. Springer Berlin Heidelberg.

[94] Zhang, B., Guan, Y., Sun, H., Liu, Q., and Kong, J. (2010). Survey of user behaviors as implicit feedback. In 2010 International Conference on Computer, Mechatronics, Control and Electronic Engineering (CMCE), volume 6, pages 345–348.

[95] Cai, Y. and Li, Q. (2010). Personalized search by tag-based user profile and resource profile in collaborative tagging systems. In Proceedings of the 19th ACM International

Conference on Information and Knowledge Management, CIKM ’10, pages 969–978, New York, NY, USA. ACM

[96] Meo, P. d., Ferrara, E., Abel, F., Aroyo, L., and Houben, G.-J. (2014). Analyzing user behavior across social sharing environments. ACM Trans. Intell. Syst. Technol., 5(1) :14 :1–14 :31.

[97] Cantador, I., Szomszor, M., Alani, H., Fernandez, M., and Castells, P. (2008). Enriching ontological user profiles with tagging history for multi-domain recommendations. In 1st International Workshop on Collective Semantics : Collective Intelligence & the Semantic Web (CISWeb 2008).

[98] Carmagnola, F., Cena, F., Cortassa, O., Gena, C., and Torre, I. (2007). Towards a tag-based user model : How can user model benefit from tags ? In Conati, C., McCoy, K., and Paliouras, G., editors, User Modeling 2007, volume 4511 of Lecture Notes in Computer Science, pages 445–449. Springer Berlin Heidelberg.

[99] Amous, I. (2002). M´ethodologies de conception d’applications hyperm´edia - Extension pour la r´eing´enierie des sites Web. Th`ese de doctorat, Universit´e Paul Sabatier, Toulouse, France.

[100] Bogers, T. and van den Bosch, A. (2009). Collaborative and Content-based Filtering for Item Recommendation on Social Bookmarking Websites. In Proceedings of the ACM RecSys’09 Workshop on Recommender Systems & the Social Web, pages 9–16, New-York, NY, USA.

[101] Zitouni, H., Berkani, L., and Nouali, O. (2012). Recommendation of learning resources and users using an aggregation-based approach. In Proceedings of the 2012 IEEE Second International Workshop on Advanced Information Systems for Enterprises, IWAISE ’12, pages 57–63, Washington, DC, USA. IEEE Computer Society.

[102] R. Rafeh and A. Bahrehmand, "An adaptive approach to dealing with unstable behaviour of users in collaborative filtering systems," Journal of Information Science, vol. 38, pp. 205-221, 2012.

[103] Abel, F., Gao, Q., Houben, G.-J., and Tao, K. (2011b). Semantic enrichment of twitter posts for user profile construction on the social web. In Proceedings of the 8th Extended Semantic Web Conference on The Semanic Web : Research and Applications - Volume Part II, ESWC’11, page 375–389, Berlin, Heidelberg. Springer-Verlag.

[104] Li, X., Guo, L., and Zhao, Y. E. (2008). Tag-based social interest discovery. In Proceedings of the 17th International Conference on World Wide Web, WWW ’08, page 675–684, New York, NY, USA. ACM.

[105] Q. Yuan, S. Zhao, L. Chen, Y. Liu, S. Ding, X. Zhang, et al., "Augmenting collaborative recommender by fusing explicit social relationships," in Workshop on Recommender Systems and the Social Web, Recsys 2009, pp. 46-56, 2009.

[106] Z. Huang, H. Chen, and D. Zeng, "Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering," ACM Transactions on Information Systems (TOIS), vol. 22, pp. 116-142, 2004.

[107] I. Konstas, V. Stathopoulos, and J. M. Jose, "On social networks and collaborative recommendation," in Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp. 195-202, 2009.

[108] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl, "An algorithmic framework for performing collaborative filtering," in Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 230-237, 1999.

[109] B. Chikhaoui, M. Chiazzaro and S. Wang, “An improved hybrid recommender system by combining predictions,” In Advanced Information Networking and Applications (WAINA), 2011 IEEE Workshops of International Conference on, pp. 644-649, 2011.

[110] M. Z. Dietmar Jannach , Alexander Felfernig,Gerhard Friedrich, Recommender Systems : an introduction. New York: Cambridge University Press, 2011.

[111] I. Fernández-Tobías, I. Cantador, and A. Bellogín, "Semantic disambiguation and contextualisation of social tags," vol. 7138 LNCS, ed, pp. 181-197, 2012.