République **A**lgérienne **D**émocratique et **P**opulaire
Ministère de l'**E**nseignement **S**upérieur et de
la **R**echerche **S**cientifique
Université Ibn Khaldoun – Tiaret –

Faculté des Sciences et de la Technologie et Sciences de la Matière

Département des Sciences et de la Technologie

MEMOIRE EN VUE DE L'OBTENTION DU DIPLÔME DE MAGISTER

SPECIALITE: Informatique
OPTION : **S**ystème d'**I**nformation et de **C**ommunication (**SIC**)

Présenté par
BENATHMANE Lalia

SUJET DU MEMOIRE :

Optimisation sémantique des systèmes de recommandation

SOUTENU LE ……2013 Devant Le Jury Composé de :

Mr. Amar BALLA	Professeur (ESI)	Alger	Président
Mr. Omar NOUALI	Directeur de recherche CERIST	Alger	Examinateur
Mr. Rachid CHALAL	Maitre de conférences A (ESI)	Alger	Examinateur
Mr. Youcef DAHMANI	Maitre de conférences A (UIK)	Tiaret	Directeur de mémoire

Année Universitaire : 2012/2013

Résumé

Le filtrage d'information est le processus permettant d'acheminer un flot continu d'informations pertinentes vers des groupes de personnes, sans qu'ils aient à exprimer explicitement ce qu'ils recherchent, depuis ces objectifs ou ces désirs définis préalablement, par le biais du principe de collaboration qui apporte des bénéfices pour tous les usagers du système en contrepartie d'un effort individuel. L'élaboration de tel système confronte à certaines limites, telles que la phase de construction initiale « démarrage à froid » et le passage à l'échelle et parfois le manque des données, et l'effet entonnoir.

Dans ce mémoire nous montrons que l'utilisation des informations sémantiques des ressources ((type, propriétés)…), peut améliorer la précision, la couverture et la qualité des systèmes de recommandation. Ainsi que la minimisation de la taille du système par la formation des communautés sémantiques qui sont aussi utilisées comme des sources des données pour recalculer les données manquantes pour avoir des prédictions plus précises .

On présente dans ce mémoire une approche de filtrage d'information hybride, qui est basée sur l'algorithme basé mémoire afin de l'améliorer et de réduire les effets des problèmes de filtrage collaboratif par l'intégration du web sémantique dans le processus de filtrage.

Mots clés : Filtrage Collaboratif, Web sémantique, Ontologie de domaine, Délégué « représentant », Similarité sémantique, Arrière plan, Système de filtrage hybride.

Abstract

The filtering of information is the process allowing to forward a continuous stream of relevant information to groups of persons, without having to express explicitly what they look for, since these objectives or these desires defined beforehand, by means of the principle of collaboration which brings profits for all the users of the system in return of an individual effort. The elaboration of such system confronts with certain limits, such as the initial construction phase "cold start" and the scale net, the lack of the data, and the funnel effect.

In this document, we show that the use of the semantic information of the resources ((type , properties)), can improve the precision, the cover and the quality of the systems of recommendation. As well as the minimization of the size of the system by the formation of the semantic communities which are also used as sources of the data to recalculate the missing data to have more precise predictions.

We present in this document an approach of hybrid information filtering, which bases on the based memory algorithm to improv and reduce the effects of the problems of collaborative filtering by the integration of semantic web in the process of filtering.

***Keywords*:** Collaborative filtering, web semantic, Ontology of domain, Delegated "representative", Semantics similarity, Background, Hybrid filtering system.

Remarciments

Tout d'abord, je remercie le bon Dieu pour m'avoir illuminée et menée jusqu'ici.

je tiens à remercier chaleureusement Pr : DAHMANI Youssef d'avoir accepté d'être mon directeur de thèse et Dr : KHAROUBI Sahraoui. Je leur suis particulièrement reconnaissante pour leur disponibilité, leur compétence, leur soutien, leurs conseils judicieux et la confiance dont ils m'ont fait part lors de la réalisation de ce travail.

Je souhaite adresser mes sincères remerciements aux membres du jury, de ma thèse, qui ont accepté la tâche délicate de rapporter ce mémoire et qui ont eu la patience de juger ce travail, et d'apporter leurs avis éclairés sur mon travail de recherche.

Mes remerciements vont également aux membres de ma famille, particulièrement mes chers parents, mes frères et sœurs

Sans oublier mon mari.

Aussi, je tiens à remercier tous mes enseignants, mes collègues de la promotion et les personnels du département l'ESI, et à toute personne qui a apporté une aide lors de la réalisation de ce mémoire.

BEN LALIA

SOMMAIRE

Liste des figures

Liste des tableaux

Introduction générale

Le filtrage collaboratif consiste à filtrer les documents du flux entrant en se basant sur les profils de chaque utilisateur. A la différence des moteurs de recherche où l'accès est actif les systèmes de filtrage présentent un accès passif à l'information, il est caractérisé par le fait que l'utilisateur reçoit des recommandations sur ce qui l'intéresse, sans l'envoi d'une requête, pour pallier le problème de la surcharge d'information, et personnaliser l'accès aux informations. Il présente aussi d'autres avantages, comme la possibilité de recommander tout type de ressources (images, vidéos etc.), la possibilité d'intégrer d'autres facteurs tels que le web sémantique. Cependant les systèmes de filtrage soulèvent des problématiques. D'abord le nombre peu important d'évaluations dans la matrice où les objets à recommander ne sont décrits que par les évaluations fournies par les utilisateurs, qui engendrent des prédictions peu pertinentes, cette problématique est nommée « **Matrice creuse** », et encore pire lors du « **démarrage à froid** » du système ou aucune information n'est disponible. Il faut aussi tenir compte du problème de « **l'effet entonnoir** » où la majorité des systèmes actuels ne permettent pas dans certains cas de prendre en compte les documents d'un nouvel axe de recherche pour les différentes communautés, et en plus la complexité de ces systèmes, les techniques doivent s'adapter aux centaines de milliers d'utilisateurs ou de ressources, en gardant un niveau de performance acceptable.

Aujourd'hui, l'adoption des systèmes de filtrage est assez importante, mais le défi est l'amélioration des pratiques et des méthodes utilisées pour rendre les systèmes plus précis et performants. Cette amélioration passe par l'amélioration de la prise en charge des problématiques liées à ces types de systèmes et par la proposition de nouvelles approches pour améliorer les fonctionnalités.

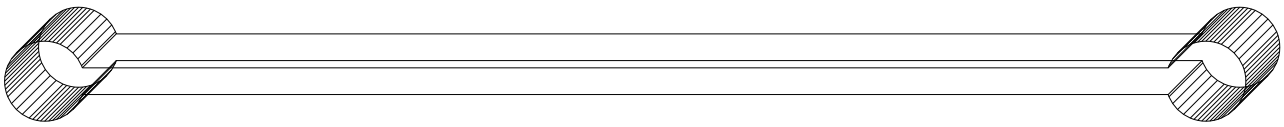
Le reste de ce manuscrit est composé comme suit :

Le « **chapitre 1** » fait le point sur l'état de l'art du filtrage d'information, nous présentons ces grandes familles « **collaboratif et basé contenu** » ensuite dans le « **chapitre 2** » nous présentons plusieurs techniques de filtrage collaboratif qui ont été explorées pour réduire les problématiques indiqués ci-dessus. Certaines sont statistiques (classification, réduction de la dimension, etc.), elles permettent d'alléger l'impact du faible nombre d'évaluations et d'améliorer la performance. D'autres visent à résoudre les problématiques de démarrage à froid, de la nouvelle entité et du nombre d'évaluations en introduisant l'aspect sémantique de différentes manières dans le processus de prédiction, et en intégrant les deux types d'approches dans des techniques hybrides.

Dans notre travail, nous nous sommes penchés sur la piste visant à introduire l'aspect sémantique, ceci en utilisant l'infrastructure du web sémantique, à cause des avantages que nous pouvons en tirer dans le « **chapitre 3** » :

Nous proposons une approche de filtrage d'information hybride au « **chapitre 4** », qui se base sur l'algorithme basé-mémoire afin de l'améliorer et de réduire les effets des problèmes présentés ci-dessus, nous proposons l'intégration du principe de l'algorithme basé modèle «le regroupement des utilisateurs similaires » et la description sémantique des documents à recommander « ontologie de domaine ». Pour cela nous avons évalué plusieurs module : La formation des communautés, Le calcule de prédiction La récupération des évaluations .En plus de la proposition d'un algorithme pour la récupération des évaluations, nous avons défini un algorithme pour le calcul de la similarité qui se base sur le nombre de recommandation de chaque item, et des résultats et des comparaisons sont donnés. Nous terminons par une conclusion et quelques perspectives.

Chapitre 1



Filtrage d’information

1. Introduction :

Le but d'utiliser un système de recherche d'information est d'accéder à l'information pertinente, adaptée aux besoins des utilisateurs, depuis des ressources hétérogènes.

Les systèmes de recherche sont divisés en deux catégories, selon la stratégie de recherche utilisée active ou passive :

- ✓ ***Accès actif :*** est utilisé par les systèmes de recherche d'informations traditionnels, caractérisé par le fait que l'utilisateur fait un effort de recherche pour trouver ce qu'il désire, l'utilisateur formule son besoin par une requête, en utilisant des mots-clés qui seront comparés avec les documents indexés dans les bases de données. Les résultats retournés aux utilisateurs contiennent souvent un grand nombre de documents non pertinents où l'utilisateur doit sélectionner manuellement les documents pertinents. Il s'agit d'une tâche pénible et ennuyeuse pour l'utilisateur
- ✓ ***Accès passif :***Caractérisé par le fait que l'utilisateur reçoit des informations ou bien des recommandations sur ce qui l'intéresse, sans l'envoi d'une requête pour pallier le problème de la surcharge d'information de la première stratégie , et personnaliser l'accès aux informations .Ces recommandations faites par un ami, une liste de diffusion (e-mails) ou bien un système de filtrage d'information.

2. Définition :

Le filtrage d'information est le processus permettant à partir d'un large volume d'informations dynamiques, d'extraire et de présenter les seuls documents intéressants un utilisateur ou un groupe d'utilisateurs ayant des centres d'intérêts relativement semblable appelés profils [Boughanem, 2001]. Ces profils (centres d'intérêt des utilisateurs) sont définis préalablement et de manière permanente, aident les utilisateurs sans expérience personnelle suffisante à faire un choix en lien avec ses goûts et ses attentes parmi les recommandations de toutes sortes.

L'objectif de l'utilisateur à utiliser ce type de système est à la fois de minimiser le temps passé à la recherche, et de lui suggérer des items[1] pertinents qu'il n'aurait pas spontanément consultés, ainsi accroître sa satisfaction globale. La figure 1.1 [Maria, 2005] donne une vision générale sur le processus de filtrage d'information, où le système fait une « «prédiction », qui

[1] Peut par exemple être une page web, un livre, un film, de la musique

s'appuie sur le « profil »d'un utilisateur et aboutit à une décision : « recommander » ou « ne pas recommander » l'information.

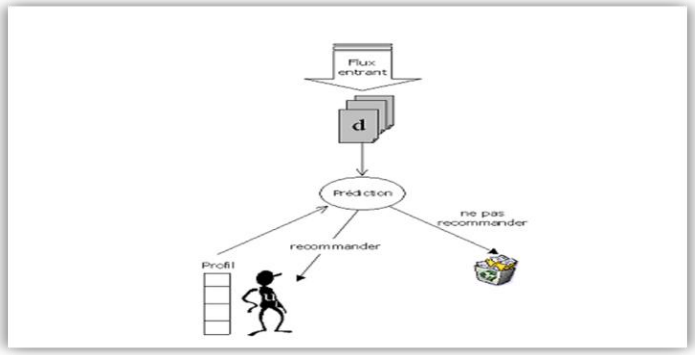


Figure 1.1 : Filtrage d'information

3. Terminologie:

Le filtrage d'information est un domaine riche qui contient une multitude de terminologies différentes, qui a besoin d'être clarifiée [Hamid, 2004] :

- ✓ ***Le filtrage d'information :*** Est alternativement défini par plusieurs noms, tels que routage, recommandation et diffusion sélective de l'information(ou SDI[2]) :
  - ***Le routage :*** Est utilisé pour indiquer qu'une liste de documents est acheminée à un ou plusieurs utilisateurs.
  - ***La recommandation :*** Consiste à exploiter les préférences d'une communauté d'utilisateurs pour prédire la préférence d'un utilisateur donné.
  - ***La SDI* :** Est utilisée pour insinuer que les profils décrivant des besoins en information sont construits manuellement, et permettent d'identifier les domaines de spécialisations des utilisateurs. On utilise de plus en plus le terme accès personnalisé à l'information. Ce terme englobe toutes les formes d'accès à l'information qui mettent l'utilisateur au centre du processus de sélection d'information.
- ✓***Les besoins en information :*** Mizzaro [Mizzaro ,1997] considère que le besoin en information passe par plusieurs phases :
  - ***La phase initiale :*** c'est-à-dire, l'état anormal de la connaissance, appelé aussi problème
  - ***Un problème*** : se transforme ensuite en un besoin en information lorsque la personne se rend compte de ce qu'elle désire comme information.
  - ***L'expression du besoin en information :*** se transforme en une requête.

[2] Selective Dissemination of new Information

Le besoin en information est souvent désigné par « centre d'intérêt». Il est parfois identifié sous le nom "topic".On retrouve également les termes profile et requête qui peuvent désigner l'expression d'un besoin selon le contexte dans lequel ils sont invoqués.

✓***Communauté:*** Une communauté est un ensemble d'utilisateurs proches les uns des autres relativement à un critère particulier.

4. Caractéristiques d'un système de filtrage :

Un système de filtrage d'information est caractérisé par les propriétés suivantes :

✓Exploite un grand volume de données, entrant transmis par des sources distantes

✓Basé principalement sur le profil de l'utilisateur ou d'un groupe d'utilisateurs

✓Diffuse que les informations en adéquation avec le profil l'utilisateur [Abbes, 1999]

✓Accès aux derniers documents arrivés [Samia, 2007]

✓Intègre l'utilisateur, avec l'évaluation des ressources recommandées, pour la mise à jour de son profil [Amokrane, 2007]

5. Fonctionnement des systèmes de filtrage :

Le principe de fonctionnement d'un système de filtrage d'information est d'acheminer des documents vers des groupes de personnes, depuis leurs *objectifs* ou leurs désirs définis préalablement qui sont relativement stables, à long terme ou périodiques. Ceci amène à des besoins réguliers d'information (exemple : être à jour sur un sujet) qui peuvent évoluer lentement au cours du temps au fur et à mesure que les conditions, objectifs et connaissances changent. La figure 1.2 illustre l'architecture de base des systèmes de filtrage d'information, telle qu'elle a été présentée par Belkin et Croft [Belkin et Croft, 1992]. La figure est caractérisée par trois blocs :

✓***Bloc 1 :*** La création de substituts (représentation) de documents :

- A l'arrivé d'un document[3], le système de filtrage associe une représentation de contenu à ce document.

✓***Bloc 2 :*** Création de profils :

- Il représente les besoins en information d'un nouvel utilisateur par des profils
- Il modifie les profils d'un ancien utilisateur.

✓***Bloc 3 :*** Le processus de comparaison et de filtrage :

[3] Les producteurs de documents distribuent leurs produits dés qu'ils sont générés

- Il compare la représentation des documents et les profils de l'utilisateur actif et décide si les documents sont pertinents ou non pour les envoyer à l'utilisateur. Cette étape peut mener dans la plupart des cas à l'amélioration des profils et des domaines d'intérêt de l'utilisateur.

Ce processus de filtrage est déclenché à l'arrivée d'un nouvel événement au Bloc 1 ou Bloc 2.

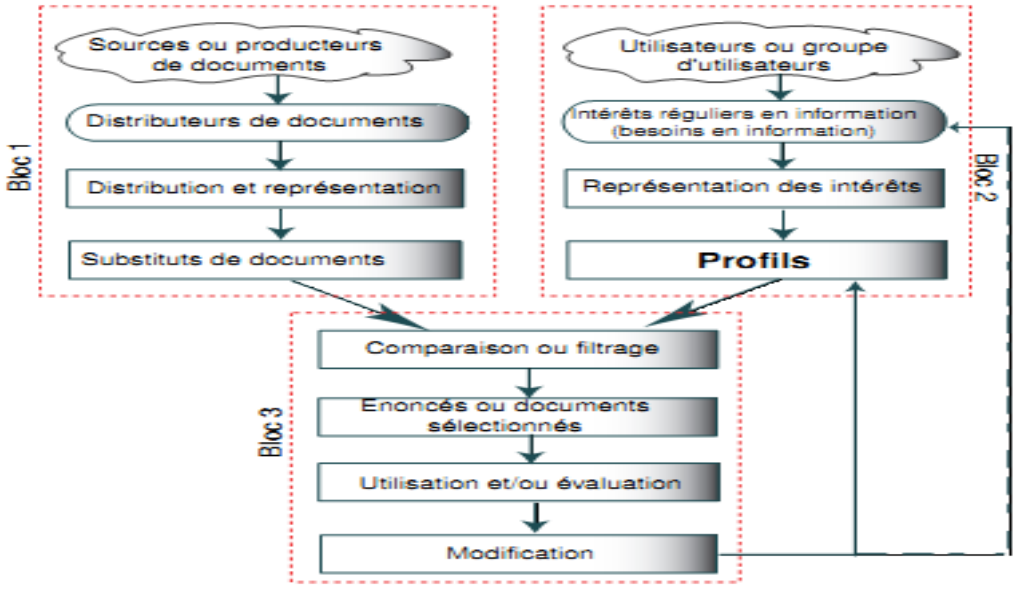


Figure 1.2 : Architecture générale d'un système de filtrage d'information [Hamid, 2004]

6. Grandes familles de filtrage d'information :

Traditionnellement, les systèmes de filtrage d'informations ont été classés en trois catégories :

- ✓ ***Filtrage cognitif :*** Basé sur le contenu.
- ✓ ***Filtrage collaboratif (social) :*** Basé sur les évaluations des utilisateurs sur les ressources.
- ✓ ***Filtrage hybrides*** : Combine les deux approches.

Cette classification dépend de la manière avec laquelle l'utilité ou la pertinence éventuelle est calculée ou estimée (au niveau du bloc 3).

6.1. Le filtrage cognitif :

Le filtrage cognitif est l'approche la plus anciennement utilisée dans le domaine du filtrage [Malone et al, 1987]. C'est un type de filtrage [Hamid ,2004] dont la décision de sélectionner un document se base uniquement sur le contenu de ce document. Chaque utilisateur du

système [An, 2006] possède un profil [4]qui décrit ses propres centres d'intérêt. À l'arrivée d'un document, le système compare la représentation du document avec le profil pour prédire la satisfaction de l'utilisateur sur ce document. Les représentations peuvent être exprimées par des mots et/ou des phrases (thèmes) plus spécifiques, dépend [Houda, 2009] de la méthode utilisée dans l'analyse du contenu des documents. Les techniques d'analyse textuelle empruntées à la recherche d'informations pour la recommandation de documents textuels (sites web, articles…), représente les profils souvent sous forme d'un vecteur de mots-clés avec des poids (1). Le poids associé à chaque mot reflète l'importance de ce terme pour l'utilisateur. Ces mots sont souvent extraits à l'aide de la mesure TF-IDF .Ce vecteur est ensuite comparé à celui du document (2), pour se faire plusieurs mesures peuvent être utilisées telles que la mesure des vecteurs cousine (3).

$$\text{Profil\_contenu}(c) = \{(t^c_j, w^c_j)\}, j = 1 \ldots k$$

$$\text{Contenu}(s) = \{(t^s_j, w^s_j)\}, i = 1 \ldots n$$

$$u(c,s) = \cos(\vec{w}_c, \vec{w}_s) = \sum_{i=1,k} \frac{w_{i,c}}{\sqrt{\sum_{i=1,k} w^2_{i,c}}} \frac{w_{i,s}}{\sqrt{\sum_{i=1,k} w^2_{i,s}}}$$

Avec : $w_{i,c}$ le poits du terme « i » dans le vecteur du profil utilisateur « c »

$w_{i,s}$ le poits du terme « i » dans le vecteur du document « s »

En plus de ces méthodes heuristiques, d'autres systèmes à base de modèles ont été proposés dans la littérature ; tels que les classificateurs bayesiens, les réseaux de neurones, les arbres de décision.

Le filtrage basé sur le contenu peut être vu comme un système de recherche d'informations dont la fonction de correspondance entre une requête et un corpus de documents joue le rôle d'un filtre permanent entre un profil (sorte de requête à long terme et évolutive) et le flot de documents entrant (sorte de corpus évolutif).Il assure les deux fonctionnalités centrales ressortent, pour un système de filtrage d'information (SFI) :

- ✓ La sélection des documents pertinents vis-à-vis du profil
- ✓ La mise à jour du profil en fonction du retour de pertinence fournis 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.

La figure 1.3 présente un processus de ce type de filtrage :

[4] Le profil peut contenir une liste des thèmes que l'utilisateur aime bien ou qu'il n'aime pas

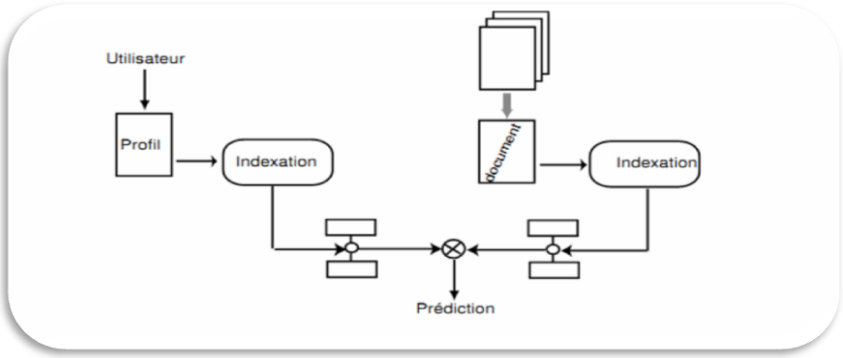


Figure 1.3 : Processus de filtrage basé sur le contenu

L'exemple le plus saillant du filtrage basé sur le contenu est le filtrage d'objets textuels (par exemple, les mails ou les pages WEB) basés sur les mots contenus dans leurs représentations textuelles. Parmi ses avantages, on peut citer :

- ✓L'utilisateur dans un tel système ne dépend absolument pas des autres
- ✓Il peut répondre aux intérêts à long terme des utilisateurs
- ✓Employant des techniques efficaces dans le domaine de l'intelligence artificielle pour la mise à jour des profils.

Par contre quelques uns de ses limites sont :

- ✓ La difficulté d'indexation les documents multimédias (image, son, vidéo)
- ✓ Basé sur le critère thématique uniquement, l'absence des autres facteurs comme la qualité scientifique, le public visé par l'auteur et l'utilisateur, etc.
- ✓ Problème du démarrage à froid : Un nouvel utilisateur du système éprouve des difficultés à exprimer son profil en spécifiant les thèmes qui l'intéressent. Ceci malgré les techniques d'apprentissage où l'utilisateur fournit des textes exemples.
- ✓ L'effet « entonnoir » : les besoins de l'utilisateur sont de plus en plus spécifiques, or les systèmes actuels ne permettent pas dans certains cas de filtrer tous les documents surtout ceux qui sont proches thématiquement mais décrits d'une manière différente ce qui l'empêche d'avoir une diversité de sujets. Même pire, un nouvel axe de recherche dans un domaine bien précis peut ne pas être pris en compte car il ne fait pas partie du profil explicite de l'utilisateur.

6.2. Le filtrage collaboratif:

À l'opposé du filtrage basé sur le contenu, le filtrage collaboratif [Malone et al, 1987] s'appuye sur la communauté des utilisateurs du système. Il a pour principe[An,2006] d'automatiser les processus sociaux par l'exploitation des évaluations que des utilisateurs ont

faite sur certains documents, afin de recommander ces mêmes documents à d'autres utilisateur , sans qu'il soit nécessaire d'analyser le contenu des documents donc cette technique peut s'appliquer dans les contextes où le contenu soit indisponible, soit difficile à analyser, et en particulier elle peut s'utiliser pour tout type de donnée: texte, image, audio et vidéo .

Par exemple, dans la figure 1.4, supposons que l'on a des communautés formées par la proximité des évaluations des utilisateurs. Le document « d » sera recommandé à l'utilisateur « u », car ce document est apprécié de la communauté G où se trouve l'utilisateur c.à.d. les utilisateurs les plus proches qu'ont émis un jugement de valeur favorable.

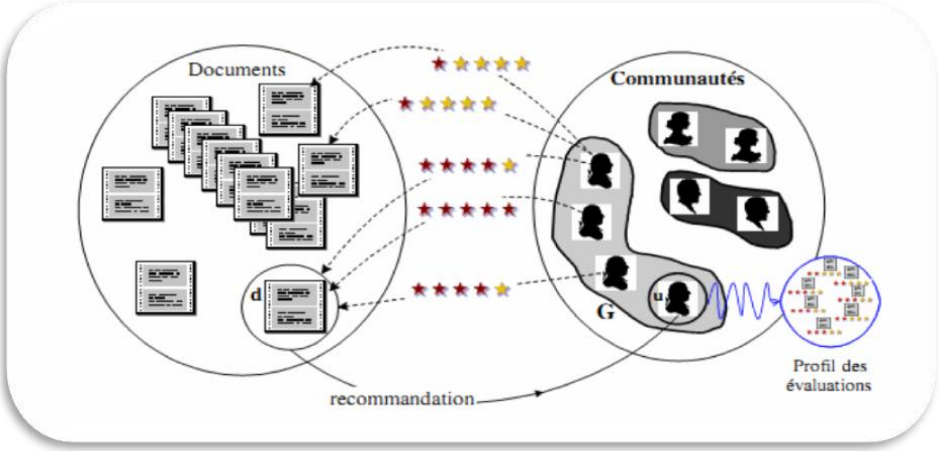


Figure 1.4 : Principe général d'un filtrage collaboratif [An, 2006]

Quant un document est arrivé [Maria, 2005], il faut avoir au moins une évaluation pour pouvoir être recommandé par ce système de filtrage. Par ailleurs, un profil se présente sous la forme d'un ensemble d'évaluations sur des documents (généralement sous la forme de votes quantitatifs) faites dans le passé par l'utilisateur .Le profil est modifié au cours du temps à partir des nouvelles évaluations que l'utilisateur réalise. Parmi ces avantages :

- ✓ Filtrer tout type d'information (images, vidéos, etc.)
- ✓ L'utilisateur capable de découvrir divers domaines intéressants
- ✓ Permet d'exprimer des autres facteurs et critères tels que : la qualité de l'information, le public visé, la zone géographique, etc. Chose qui n'est pas possible dans un système de filtrage thématique
- ✓ L'effet entonnoir est atténué, car tout document évalué par une personne peut être recommandé. L'utilisateur peut bénéficier de nouveaux axes de recherche auxquels il n'a jamais pensé.

Cependant, on peut citer quelques uns de ses limites:

- ✓ ***Démarrage à froid :*** Ce phénomène se produit au début d'utilisation de ce système, à l'arrivée d'un nouvel utilisateur, lorsqu'il s'inscrit pour utiliser le système, sa communauté est encore inconnue, ce qui conduit à l'impossibilité de fournir des recommandations pertinentes.
- ✓ ***Masse critique :*** Afin de former de meilleures communautés, le système exige un nombre suffisant d'évaluations en commun (même opinion) entre les utilisateurs pour les comparer entre eux. Et pourtant, vu la taille énorme de l'ensemble des documents, achats, etc. dans les systèmes, le nombre des évaluations en commun entre utilisateurs risque d'être faible.
- ✓ ***Rapport coût/bénéfice :*** Pour qu'un système de filtrage collaboratif soit utile il faut que l'utilisateur effectue un nombre minimal d'évaluations, l'utilisateur peut abandonner s'il n'est pas convaincu du résultat à court ou à moyen terme.
- ✓ ***Expression du besoin:*** Un système de filtrage par défaut recommande les ressources pouvant intéresser l'utilisateur, mais ne prend pas en charge le contexte dans lequel est l'utilisateur. Donc l'utilisateur peut recevoir des recommandations qui ne correspondent pas à leur contexte.

•La comparaison de l'approche cognitive et collaborative :

La figure 1.5 présente une brève comparaison entre les SFI collaboratif et cognitif :

	Filtrage basé sur le contenu sémantique	**Filtrage collaboratif**
Amorçage (démarrage de l'exploitation du système)	Le filtrage peut commencer après l'établissement du profil	Exige une base de données substantielle et plusieurs évaluations de l'utilisateur avant d'être utilisable
Qualité de l'information (lisibilité, fiabilité, nouveauté, etc.)	La qualité de l'information n'est pas connue	La qualité de l'information est connue *via* des évaluations d'utilisateurs
Contexte de l'information (domaine d'intérêt)	L'identification du domaine se fait généralement par la co-occurrence des termes dans chaque document	L'identification du domaine se fait par la différence des domaines d'intérêt des utilisateurs
Effet « entonnoir »	Le système ne suggère que des documents dont le thème a déjà été évoqué explicitement	Le système peut suggérer des documents sans rapport explicite avec les thèmes déjà évoqués

Figure 1.5 : Comparaison de l'approche cognitive et collaborative

6.3. Filtrage hybride :

Le principe d'hybridation [An, 2006] s'effectue en deux phases :

- ✓Appliquer séparément le filtrage collaboratif et autres techniques de filtrage pour générer des recommandations candidates
- ✓Combiner ces ensembles de recommandations préliminaires selon certaines méthodes telles que la pondération, la commutation, etc.…, afin de produire les recommandations finales pour les utilisateurs.

Plus généralement, les systèmes hybrides gèrent des profils d'utilisateurs orientés contenus, et la comparaison entre ces profils donne lieu à la formation de communautés d'utilisateurs permettant le filtrage collaboratif.

6.4. Autres types de filtrage :

En plus de ces 3 types de basse, le filtrage d'information a connu quelques variantes combinant des critères supplémentaires afin de l'améliorer .On peut à titre d'exemple citer :

✓**Filtrage actif :**

Les limitations du filtrage collaboratif peuvent trouver des solutions dans le filtrage dit actif qui contribue à réduire le démarrage à froid par la possibilité offerte aux utilisateurs de la communauté de se recommander mutuellement des documents. Lorsqu' un utilisateur trouve des documents plus ou moins intéressants pour certains autres utilisateurs qu'il connaît, il peut les leur recommander [An et al, 2004].

✓**Filtrage différé :**

Le filtrage différé consiste à stocker les documents dès leur arrivée et à filtrer chaque ensemble de documents stockés disponibles à la fin de chaque période déterminée par le système de filtrage, il diffère de la recherche d'informations par le fait que les documents arrivent séquentiellement à travers le temps [Boughanem, 2001].

✓**Filtrage adaptatif :**

Un système de filtrage adaptatif, ou système de recommandation, doit faire parvenir continuellement des informations pertinentes aux utilisateurs tout en s'adaptant en permanence à leurs besoins d'information [An et al, 2005]. Le système peut tirer parti de la pertinence ou de la non-pertinence des documents sélectionnés et filtrés pour l'utilisateur pour améliorer ses performances au cours du temps [Bisiaux ,2003], le système commence avec seulement un profil utilisateur et un petit nombre de documents pertinents, il doit procéder au filtrage de documents sans aucune autre information a priori.

Chaque document filtré est immédiatement jugé sur sa pertinence .Cette information sera alors exploitée par le système pour mettre à jour le profil et réadapter sa fonction du filtrage.

7. Evaluation des performances des systèmes de filtrage:

Pour évaluer les systèmes de filtrage d'information, il y a plusieurs métriques, classifier en deux catégories les métriques de classification, et prédictives :

7.1. Les métriques prédictives :

Pour évaluer les algorithmes de filtrage collaboratif « P », on utilise généralement une technique statistique appelée « validation croisée » (*cross-validation*).Elle consiste à séparer les données disponibles en sous-ensembles. La première partie sert à faire la prédiction et la deuxième, à valider (Montrer la généralité du système de filtrage proposée c.à.d. attendre tous les objectifs déclarer) l'algorithme. En pratique, cela se passe comme suit :

- ✓ Nous choisissons aléatoirement un certain nombre d'individus, par exemple, la moitié de ceux-ci. Ces individus sont représentés par l'algorithme P dans les calculs (validation).
- ✓ Les autres individus constituent un ensemble « test ».

Les individus de l'ensemble « test » sont alors pris un à un. Pour chaque individu « x », on note leur évaluation réelle par « r_i » à l'article « i », et on fournit par l'algorithme P une évaluation prédite « p_i ».

Ensuite on mesure le pourcentage (%) d'erreur faite en prédisant la note accordée par les individus de « test » aux articles « i » par la fonction « MAE » :

$$MAE = \frac{\sum_{i=1}^{N} |p_i - r_i|}{N}$$

Une autre mesure populaire est l'erreur de moyenne carrée de racine (RMSE). Cette dernière mesure tend à pénaliser davantage les algorithmes qui font parfois des erreurs importantes puisque la mise au carré amplifier la contribution des grandes erreurs, Elle convient aux situations où les petites erreurs de prévision ne sont pas très importantes.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (pi - ri)^2}{N}}$$

Ces mesures sont incomplètes pour plusieurs raisons :

- ✓ ***Le temps de calcul :*** parfois est un facteur déterminant des algorithmes qui n'est pas traité par ces métriques.

- ✓***La précision des évaluations prédites par le système :*** Si pour un utilisateur, leurs goûts et leurs opinions sont en constante valeur, les évaluations prédites par le système à tester restent constante. Malgré qu'il est probable que si nous demandions à des utilisateurs de noter à nouveau les mêmes articles de test, leur réponse changerait.
- ✓***La prédiction de l'évaluation réelle des utilisateurs*** : Si on demande au système de proposer à des utilisateurs les 10 meilleurs items, il choisira les items correspondant aux 10 notes les plus importantes qui ont été prédites (déterminer par un seuil), il ne prédira pas les items qu'accorderaient les utilisateurs.

Par ailleurs, il existe des solutions de rechange pour mesurer la qualité d'une recommandation « *les métriques de classification* ».

7.2. Les métriques de classification:

Dans la pratique, il faut savoir tenir compte de la perception des utilisateurs. Même si vous pouviez lire dans leurs pensées, il n'est pas évident que le fait de leur faire dire ce qu'ils veulent vraiment mène à un taux de satisfaction élevé. Pour mesurer la qualité d'une recommandation, plusieurs mesures, ont été utilisées:

Soit le tableau 1 qui correspond aux résultats d'une expérience statistique classique :

Actuel \ Prévu	Négative (ne sont pas prédits)	Positive (sont prédit)
Négative	a	b
Positive	c	d

Nous avons les métriques suivantes :

- ✓ *l'exactitude* **:** Peut être vue comme la capacité du système à sélectionner tous les documents pertinents, elle peut aussi être vue comme la capacité du système à prédire des évaluations pour tous les documents présents dans le système, ces documents n'étant pas nécessairement pertinents :

$$\text{Exactitude} = \frac{\text{recommandations correctes}}{\text{recommandations possibles}} = \frac{a+d}{a+b+c+d}$$

- ✓ *MAD* **:** Une mesure commune d'erreur est l'erreur absolue moyenne (MAE, également appelé déviation absolue moyenne ou MAD) :

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\epsilon_i| = \frac{b+c}{a+b+c+d'}$$

Où ϵ_i :l'erreur absolus de chaque item « i ». ϵ_i Peut être un zéro si :

a=d=0 ou b=c=0

- ✓ *Rappel/précision*: On retrouve celles empruntées à la recherche d'information : la « Précision » et le « Rappel » (Salton and McGill, 1983; van Rijsbergen, 1979). Elles partitionnent l'ensemble des documents recommandés, en deux catégories : les documents pertinents et les documents non pertinents, d'où leur nom. Ainsi la précision mesure la capacité du système à rejeter tous les documents non pertinents, le rappel quant à lui mesure la capacité du système à retrouver tous les documents pertinents :

$$\text{Précision} = \frac{\text{documents positives recommandés}}{\text{documents recommandés}} = \frac{d}{b+d}$$

$$\text{Rappel} = \frac{\text{documents positives recommandés}}{\text{documents positives}} = \frac{d}{c+d}$$

Mais ces deux mesures d'évaluation [Samia, 2007] ne sont pas applicables dans le cas de filtrage parce qu'il est quasiment impossible de calculer le rappel, car l'utilisateur ne dispose pas des informations pertinentes non sélectionnées (système dynamique), en plus de ça si on considère deux systèmes de filtrage différents pour un même flux de documents :

- ✓ *Le premier système :* sélectionne une liste de 100 documents non pertinents et zéro document pertinent,
- ✓ *Le deuxième système :* sélectionne un document non pertinent et aucun document pertinent.

Si nous calculons les valeurs de précision et de rappel, on peut constater qu'elles sont nulles ; néanmoins, dans la pratique, le deuxième système est plus performant que le premier, puisque l'utilisateur ne perd pas de temps à lire des documents qui ne l'intéressent pas. Ceci rend les mesures de précision et de rappel inadéquates, car elles ne permettent pas de différencier entre les systèmes.

Pour remédier aux insuffisances de ces mesures, des autres mesures on été proposées tell que la fonction de :

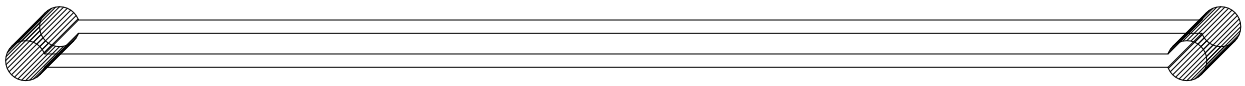
$$\text{F} - \text{beta} = \frac{0.25 * \text{d}}{\text{d} + \text{b} + 0.25 * (\text{d} + \text{b})}$$

8. Conclusion :

Le filtrage de l'information est le processus permettant à partir d'un large volume d'informations d'extraire et de présenter les seuls documents intéressants à un utilisateur ou à un groupe d'utilisateurs ayant des intérêts relativement semblables, appelé les profils. Traditionnellement, des systèmes de filtrage de l'information ont été classifiés dans deux catégories : filtrage cognitif (filtrage basé-contenu), et filtrage collaboratif. Ce dernier est le plus populaire ; il apporte des avantages à tous les utilisateurs du système, tandis qu'un effort individuel est exigé pour le filtrage cognitif.

Dans la suite, nous présentons plus en plusieurs détails le filtrage collaboratif qui est Le noyau de la problématique de cette thèse.

Chapitre 2



Filtrage collaboratif

1. Introduction :

Le travail collaboratif suscite de jour en jour l'intérêt de centre socio-économique et même le monde de l'université et la recherche scientifique, il n'y a qu'à voir le nombre croissant de plateformes et d'outils qui sont développés dans ce domaine. Le filtrage collaboratif vient renforcer l'idée que les personnes à la recherche d'informations devraient pouvoir se servir de ce que d'autres ont déjà trouvé et évalué.

Ce chapitre a pour objectif de présenter en détail les systèmes de filtrage collaboratif.

2. Définition :

Le filtrage collaboratif consiste à filtrer les documents du flux entrant en se basant sur l'opinion que, chaque utilisateur de la communauté, a porté dessus. Tout document qu'il l'aura alors jugé intéressant, sera diffusé à l'ensemble des utilisateurs ayant eu des opinions similaires par le passé. Le système emploie des méthodes statistiques pour faire des prévisions selon les intérêts des utilisateurs, ces prévisions vont servir pour proposer un document à un utilisateur selon la corrélation avec le profil des utilisateurs de son voisinage. L'utilisateur après avoir reçu le document, l'évalue à son tour en lui attribuant un score en fonction de son appréciation sur sa pertinence, le système réajustera alors automatiquement et en conséquence le profil de l'utilisateur [Samia, 2007].

Différents termes ont été utilisés pour décrire le filtrage collaboratif. Initialement, il est désigné par le filtrage social. Certains auteurs montrent que dans un système automatisé les recommandations ne peuvent pas nécessairement collaborer avec les destinataires et les recommandations peuvent être inconnues entre elles, ainsi d'autres auteurs préfèrent utiliser le terme système de recommandation [Hamid, 2004].

3. Architecture générale :

L'architecture générale d'un système de filtrage collaboratif s'articule autour de deux entités centrales : les communautés des utilisateurs, et la liste d'évaluations des documents par communauté qui sont utilisés pour émettre des recommandations, la figure 2.1 le démontre. Chaque utilisateur appartenant à une communauté spécifique, pour qu'un document « D » est envoyé à un utilisateur « A » de la communauté « B » il faut que « D » accumule un certain niveau d'approbation depuis des membres de « B ».

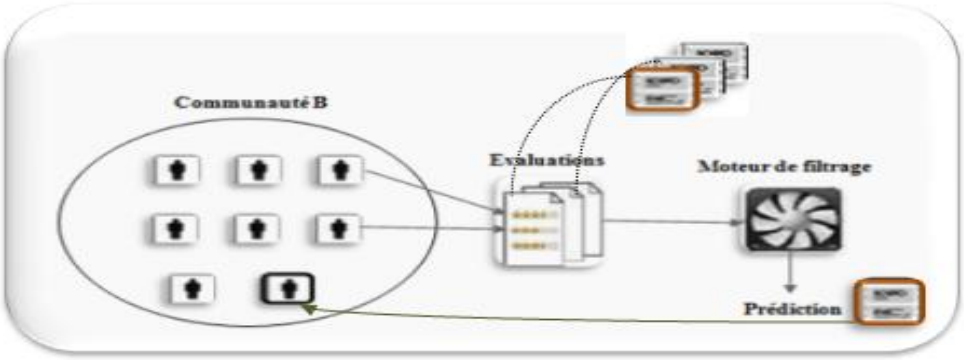


Figure 2.1 :L'architecture générale d'un système de filtrage collaboratif

4. Processus de fonctionnement :

On peut également le voir dans la figure 2.2 [An, 2006], Il y a trois processus principaux dans un système de filtrage collaboratif qui sont:

- ✓ ***La formation des communautés :*** Exécuté par le système à chaque mise à jour des profils ou à l'arrive d'un nouvel utilisateur.
- ✓ ***La production des recommandations*** : Exécuté par le système à l'arrive d'une nouvelle information (reformation des communautés...).
- ✓ ***L'évaluation des recommandations*** : Exécuté par l'utilisateur, à la réception d'une recommandation.

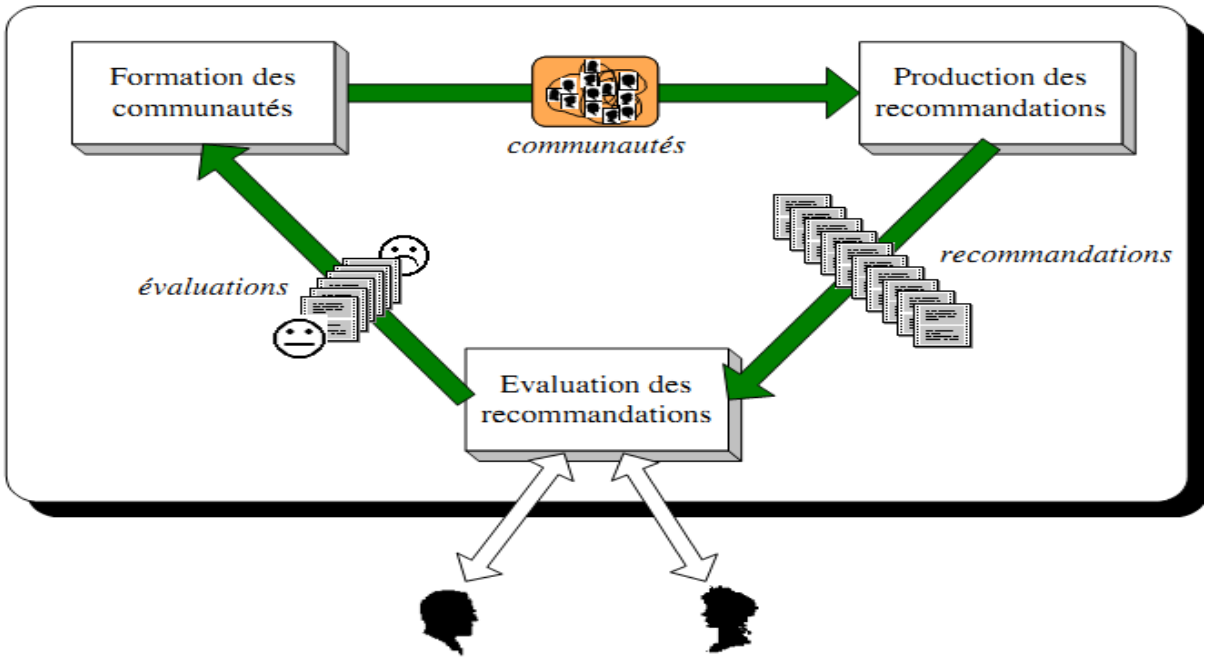


Figure 2.2 : Les trois processus principaux du filtrage collaboratif [An, 2006]

4.1. La formation des communautés :

Avant de présenter les différentes approches utilisées, nous commençons par donner une définition claire de la notion de communauté :

4.1.1. Définition :

La notion de communauté dans un système de filtrage collaboratif est définie comme le regroupement des utilisateurs en fonction de l'historique de leurs évaluations, afin que le système calcule des recommandations [An ,2006].

4.1.2. Approches de formation :

La formation des communautés est le noyau d'un système de filtrage collaboratif, aide dans la production de recommandations, consiste à regrouper les utilisateurs qui ont des propriétés communes, pour cela plusieurs approches sont employées, et qui peuvent être classifier en deux catégories suivant le type d'information utilisé pour la formation des communautés :

- ✓***Classe A :*** Regroupe les approches qui sont basées sur les informations explicites disponibles au niveau des profils des utilisateurs, l'approche la plus populaire est celle de la proximité des évaluations des utilisateurs. C'est l'approche la plus utilisée pour la formation des communautés, pour ce faire, on applique souvent la méthode des voisins les plus proches en utilisant un seuil pour le niveau de proximité ou un seuil pour la taille maximale de la communauté .La méthode des voisins les plus proches est actuellement la plus populaire. Les voisins sont identifiés à partir d'une évaluation de la similarité des appréciations sur les items communs à l'utilisateur actif et les autres utilisateurs .Elle se réalise en général en deux étapes [An ,2006]:

- Mesurer la (dis) similarité entre l'utilisateur actif « u » et les autres, et
- Sélectionner les meilleurs voisins en fonction de la (dis) similarité entre l'utilisateur « u » et les autres calculée dans l'étape précédente.

Etape 1 : ***Mesurer la (dis) similarité :***

Elle est basée sur la matrice des évaluations V_{mxn} illustrée dans la figure 2.3 :

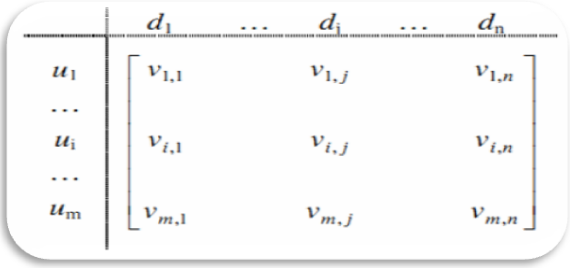


Figure 2.3 : Matrice d'évaluations [An, 2006]

Où les lignes représentent les utilisateurs U = {u_1,…,u_m } ,et les colonnes constituent les documents D={d_1,…,d_n} .Chaque case de la matrice correspond à l'historique des évaluations d'un utilisateur sur un document sous forme de notes v, identifié par :

$V_{i,j}$:l'évaluation de l'utilisateur « u_i »sur le document « d_j»

Plusieurs mesures ont été exploitées dans le but d'évaluer les similarités d'appréciations entre utilisateurs et identifier les utilisateurs voisins (les plus proches).Parmi ces mesures nous pouvons citer : le coefficient de corrélation de Pearson [Herlocker , 1999], la mesure basée sur le cosinus [Sarwar et al.,2000b], la corrélation de Spearman [Resnick , 1994]

Les mesures les plus populaires sont le coefficient de corrélation de Pearson et la mesure basée sur le cosinus. Cette popularité est liée à leur contribution à la performance des systèmes de recommandation [Anand et Mobasher, 2005].

Nous décrirons ces deux mesures ci-dessous. La matrice V_{mxn} est traitée par paires de lignes (v_u, v_i) entre l'utilisateur actif « u » et les autres utilisateurs « i ».

✓**Le coefficient de corrélation de Pearson :** cette mesure est calculée par :

$$\text{corrélation}(v_u, v_i) = \frac{\sum_i (v_{u,j} - \overline{v_u})(v_{i,j} - \overline{v_i})}{\sqrt{\sum_i (v_{u,j} - \bar{v}_u)^2 . \sum_i (v_{i,j} - \bar{v}_i)^2}}$$

Où $\overline{v_i}$: L'évaluation moyenne des items pour l'utilisateur i $\overline{v_i} = \frac{1}{|I_i|}\sum_{j \in I_i} v_{i,j}$

La figure 2.4 [DO Minh Chau, 2007] présente la signification des valeurs possible de la corrélation de Pearson :

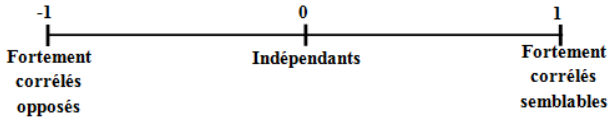


Figure 2.4 : Les valeurs possibles de la corrélation Pearson

✓**La mesure basée sur le cosinus :** Cette mesure adaptée pour l'évaluation de la similarité entre deux utilisateurs « a » et « i » en calculant le cosinus de l'angle entre les vecteurs correspondant à ces deux utilisateurs [Breese , 1998], en prenant en considération les items co-notés. La valeur calculée par la mesure cosinus est comprise entre 0 et 1.

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

L'inconvénient des mesures Pearson et cosinus [Ilham, 2010], est que le calcul des similarités devient non fiable voire impossible, lorsque le système dispose de peu d'items co-notés entre utilisateurs. Afin de résoudre ce problème, certaines extensions ont été proposées par [Breese, 1998], telle que :

- ***La note par défaut :*** Consistant à attribuer une valeur par défaut à une note manquante. Mais l'enjeu à ce niveau est de savoir quelle valeur par défaut choisir (appréciation positive, négative ou bien neutre) et d'évaluer son impact sur le calcul des similarités.
- ***L'amplification de cas :*** Permettant de transformer les similarités en amplifiant les valeurs proches de 1 et en pénalisant celles qui sont proches de 0, dans le but d'attribuer un poids important aux voisins fortement similaires à l'utilisateur actif.
- ***La fréquence inverse utilisateur*** : Inspirée de la méthode IDF [5][Salton, 1983]. L'hypothèse est que les items appréciés par un grand nombre d'utilisateurs sont moins pertinents pour le calcul des similarités entre les utilisateurs, comparés à ceux qui sont appréciés par un nombre restreint d'utilisateurs. Ainsi, chaque note est transformée en la multipliant par la fréquence inverse utilisateur qui est équivalente à log (n/nk), où « n » étant le nombre total des utilisateurs et «nk» le nombre d'utilisateurs ayant noté l'item « k » [Prem, 2010].

À la fin de l'étape 1 pour chaque utilisateur « u », on obtient une table ordonnée (Si≤ Sj, i ≤ j) par (dis) similarité illustrée dans la figure 2.5 :

Utilisateur	s_i = *dissimilarité*(V_u , V_i)
u_1	s_1
---	---
u_d	s_d
---	---
u_K	s_K
---	---
---	---
u_t	s_t

Figure 2.5 : Table ordonnée des similarités entre l'utilisateur « u » et tous les autres

[5] Inverse Document Frequency

Etape 2 : *Sélectionner les meilleurs voisins:*

Plusieurs stratégies sont définies pour sélectionner les meilleurs voisins d'un utilisateur actif « a ».On dénombre parmi ces stratégies :

✓***La détermination d'un seuil de similarité*** [Breese , 1998] [Shardanand et Maes,1995] : Il s'agit de sélectionner les plus proches voisins qui sont corrélés avec l'utilisateur actif à partir d'un seuil de similarité préétabli (figure 2.6) :

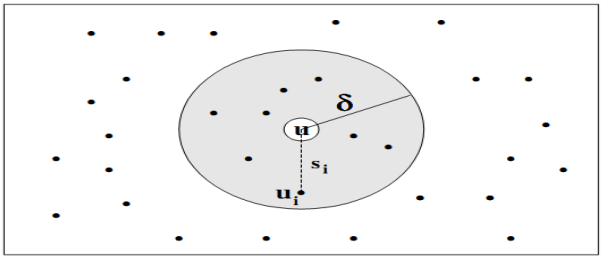


Figure 2.6 : Illustration de sélection des voisins les plus proches par le seuil δ

✓***La sélection de la taille du meilleur voisinage*** [Herlocker, 1999] : cette stratégie permet de sélectionner les voisins les plus proches (20, 50 ou 100 meilleurs voisins par exemple). On fixe un seuil K pour la taille maximale de l'ensemble des voisins (K voisins les plus proches).

✓***La détermination d'un seuil pour les items co-notés*** [Viappiani et al., 2006] : cette stratégie consiste à filtrer les plus proches voisins en fonction du nombre d'items co-notés avec l'utilisateur actif.

Au niveau des trois stratégies, les seuils choisis ne doivent pas avoir des valeurs extrêmes (ni trop élevées, ni trop faibles). En effet, par exemple, si la valeur du seuil de similarité est trop faible, cela peut engendrer de mauvaises prédictions quand l'utilisateur actif est corrélé avec de nombreux utilisateurs. De la même façon, si le seuil est très élevé, cela peut affecter la qualité des prédictions et la couverture[6], quand l'utilisateur actif est faiblement corrélé avec les autres utilisateurs. En effet, dans ce cas, le système ne dispose que de peu de voisins pour pouvoir générer les prédictions [Ilham , 2010].

✓***Classe B :*** Regroupe les approches qui sont basés sur les informations implicites qu'on peut capturer et découvrir via les tâches effectuées par les utilisateurs sur le web (historique de navigation, historique de recherche…), l'approche la plus populaire est celle des réseaux. Cette approche est basée dans son fonctionnement sur les informations qui

[6] la capacité du système à générer des prédictions

peuvent être découvertes à partir de données véhiculant de façon implicite au sein d'un réseau social. En général, leur processus se compose de trois phases [An ,2006]:

- ✓***Phase1 :*** Collecter et fouiller des données transactionnelles, par exemple communication, messages, évaluations, etc.
- ✓***Phase2 :*** Reconnaître et modéliser des intérêts souvent implicites, et induire les communautés existantes
- ✓***Phase3 :*** Explorer et exploiter des communautés.

4.1.3. Les problématiques de la gestion des communautés :

La phase de formation des communautés c'est la base du processus de filtrage collaboratif a une influence sur la qualité des recommandations envoyées aux utilisateurs, à chaque fois que ce processus est déclenché en raison d'un nouvel utilisateur qui demande de s'intégrer dans le système ou les profils d'un utilisateur sont changés, trois problématiques sont souvent à aborder (figure 2.7) :

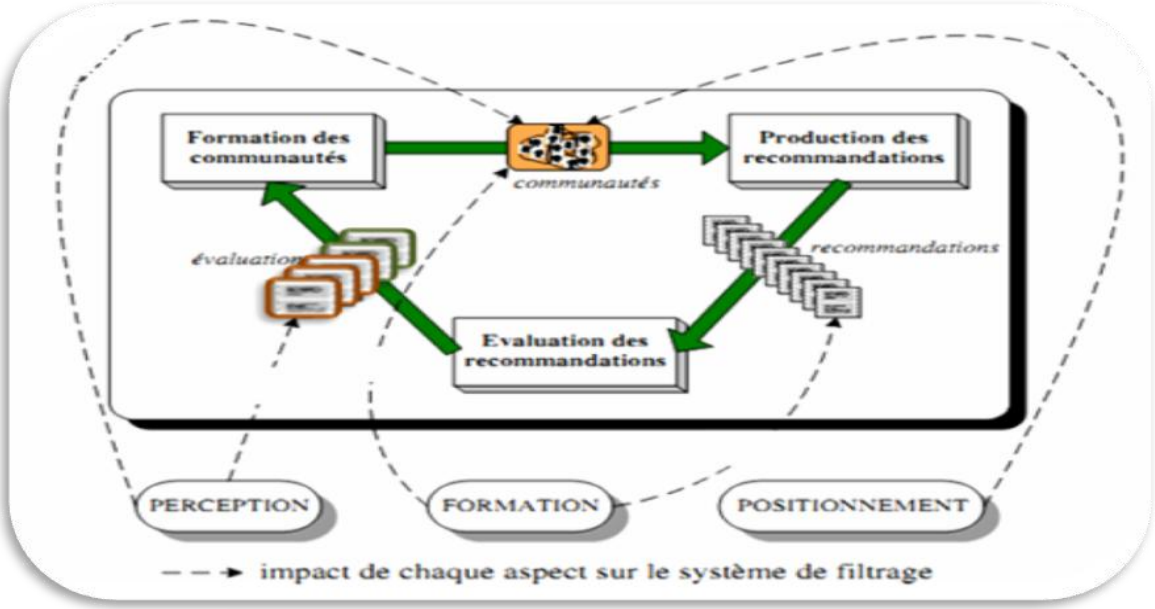


Figure 2.7 : Les problématiques de la gestion des communautés [An, 2006]

a. *Perception des communautés pour les utilisateurs* :

C'est ce que montre la figure 2.7, à la fin de chaque processus de formation, tous les utilisateurs sont intégrés dans une communauté sans intervention de façon automatique, implicite et invisible. Ce qui conduit à limiter les performances du système de filtrage collaboratif [Goldberg ,1992] [Resnick ,1994], [Breese ,1998],[Herlocker,1999] , et élimine le droit des utilisateurs de percer les communautés c.-à-d. obtenir une vision globale sur les autres participants pour comprendre ce qui se cache derrière les recommandations envoyées au cours du temps par le système .Cette situation peut rapporter des dissensions comme :

- ✓ Un utilisateur n'est pas d'accord pour les recommandations envoyées au cours du temps par le système
- ✓ Un utilisateur souhaite choisir lui-même une communauté
- ✓ Un utilisateur souhaite changer la communauté choisie par le système.

La figure 2.8 illustre les différentes solutions proposées pour résoudre ces problèmes, afin d'améliorer la confiance des utilisateurs dans les recommandations générées à partir de ces communautés.

Le niveau de perception	Les solutions	Proposée par	Idée générale
Les communautés invisibles comme un facteur interne de recommandation	Communautés dans l'explication des recommandations	[Herlocker ,2000] [Herlocker et al,2000]	Focalise sur la capacité d'explication des recommandations aux utilisateurs, afin de renforcer leur confiance et par conséquent les encourager à évaluer les recommandations reçues par l'utilisation d'un modèle de boîte blanche
		[Carenini ,2003] [Schafer ,2002] [Swearingen ,2002]	La construction d'interfaces conviviales encourageant les utilisateurs à fournir de plus en plus d'évaluations
la perception partielle des communautés	Filtrage collaboratif actif	[Maltz,1995]	Si un utilisateur trouve des documents plus ou moins intéressants pour certains autres utilisateurs qu'il connaît, il peut les leur recommander.
	Plateforme COCoFil	[Denos ,2004]	Une plateforme de filtrage collaboratif particulièrement orientée vers la communauté. Elle intègre des fonctionnalités destinées à mieux exploiter la notion de communauté d'utilisateurs. Toutes ces fonctionnalités ont pour but de permettre aux utilisateurs d'intervenir dans le processus de recommandation directement et de se découvrir entre membres d'une même communauté.

Figure 2.8 : Perception des communautés pour les utilisateurs

b. ***Formation monocritère des communautés:***

les communautés sont en général formées par la proximité des évaluations passées des utilisateurs, en appliquant souvent le principe de la méthode des voisins les plus proches, afin que chaque utilisateur soit intégré à une communauté spécifique ,où l'utilisateur reçoive uniquement les recommandations calculées à partir des évaluations des autres membres de sa communauté , quoiqu'on puisse former autour de lui autant de communautés qu'elle le souhaite : la communauté de ses proches, celle de ses collègues de travail ou plus généralement toute communauté de personnes avec lesquelles elle partage un centre d'intérêt. C'est le problème de la formation monocritère des communautés par l'historique des

évaluations dans les systèmes de filtrage collaboratif classiques, qui limite l'enrichissement et la diversification des recommandations pour les utilisateurs.

c. *Positionnement des utilisateurs dans les communautés :*

Le positionnement des utilisateurs dans les communautés dépend fondamentalement de la qualité des valeurs données pour chaque utilisateur sur chaque critère. l'absence de valeur pour un ou plusieurs critères conduit à une difficulté de positionner les utilisateurs dans les communautés ,et de choisir la bonne communauté .Les deux grands problèmes qui peuvent arriver sont le démarrage à froid et le matrice creuse:

✓ ***Le démarrage à froid pour un nouvel utilisateur :***

Le démarrage à froid est le phénomène qui se produit en début d'utilisation du système, dans des situations critiques où le système manque de données pour procéder à un filtrage personnalisé de bonne qualité,il est ainsi incapable de recommander des documents à des utilisateurs tant qu'il ne dispose pas de suffisamment d'informations sur leurs préférences et centres d'intérêts.Ce problème est généralement traité par l'utilisation de l'approche des recommandations exploratoires [Nguyen,1998],[Kohrs,2001], où la construction de l'ensemble exploratoire joue un rôle capital dans le succès du système. La figure 2.9 présente les méthodes de base pour sélectionner les recommandations exploratoires :

Les méthodes	Proposé par	Idée de base
Au hasard	[Movielens]	Les recommandations sont choisies au hasard par le système ou par l'utilisateur lui-même
Choix personnel	[Nguyen ,1998]	Les recommandations sont choisies par l'utilisateur lui-même .à l'inscription, le nouvel utilisateur peut citer ce qu'il aime en particulier et /ou qu'il n'aime pas
Popularité	[Movielens]	Le système sélectionne les documents les plus récents ou les plus évalués dans la passée
Entropie	[Kohrs, 2001]	Le système préfère les documents informatifs qui permettent de séparer les utilisateurs, plutôt qui sont appréciés par la plupart des gens

Figure 2.9 : Les méthodes de base pour générer les recommandations

Le système Entrée [Burke, 2002] utilise une méthode de combinaison du filtrage collaboratif avec le filtrage basé sur le contenu, afin de résoudre le problème du démarrage a froid, où le système demande à l'utilisateur de définir ses centres d'intérêts, en termes de contenu, à partir d'une liste de termes et/ou d'exemples décrivant au mieux ses centres d'intérêts. Ou bien on utilise le filtrage collaboratif actif [Maltz, 1995] ,où ce dernier donne aux utilisateurs les possibilités de former eux-mêmes des communautés par la connaissance de personnes, collègues ou amis .

✓ ***Matrice creuse:***

Dans le filtrage collaboratif les objets à recommander ne sont décrits que par les évaluations fournies par les utilisateurs. Mais dans la réalité, il est impossible d'obliger les utilisateurs à évaluer les recommandations, où on peut avoir un utilisateur qui à un faible nombre de

ressources communément évaluées , dans ce cas on ne peut pas s'avoir si l'utilisateur est dans la méilleure communauté ou pas . Des solutions sont proposées pour résoudre ce problème telles que les techniques à base d'agents. [Park, 2006] propose l'utilisation des agents comme moyen d'évaluation automatique.

4.2. La production des recommandations :

On peut définir ce processus comme une fonction booléenne a deux paramètres, document et utilisateur. un document « d » sera recommandé à l'utilisateur « u » qui est intégré dans la communauté « G », si et selement si ce document est apprécié de la communauté « G » c.à.d. les utilisateurs les plus proches de « u » émettent un jugement de valeur favorable. Ce processus est déclenché généralement pour deux raisons : à l'arrivée d'un document« new item » ou un nouvel utilisateur est intégré dans une communauté « new user ».

4.2.1. La production des recommandations cas de nouveau document:

C'est un problème pour le filtrage collaboratif, ce document n'est pas encore évalué, et les objets à recommander ne sont décrits que par les évaluations fournies par les utilisateurs, ce problème nommé « *le démarrage à froid pour un nouveau document (new item)* ».Ce problème est généralement traité en combinant une approche de filtrage basée sur le contenu avec le filtrage collaboratif « approche hybride », par exemple en utilisant la similarité, au niveau du contenu, entre documents pour estimer la satisfaction des utilisateurs sur le nouveau document en fonction de leurs évaluations sur certains documents assez proches [Schein, 2001];ou en introduisant des agents intelligents qui évaluent les documents automatiquement [Good ,1999].Une fois ce document est envoyé et évalué par un utilisateur dans une communauté, on peut voir les autres membres de la communauté comme des nouveaux utilisateurs .

4.2.2. La production des recommandations cas de nouvel utilisateur:

Pour la production des recommandations à un utilisateur actif « u », le système prédit l'intérêt de chaque document évalué par les membres de communauté de « u », quand il dépasse un certain seuil, le système recommande le document à l'utilisateur actif. Les techniques de calcul de prédiction peuvent être classées en trois grandes catégories : les algorithmes basés «mémoire», les algorithmes basés «modèle» et les algorithmes basés sur un «apprentissage automatique».

4.2.2.1. Algorithmes basés « mémoire » :

Ce type d'algorithmes est connu aussi sous le nom d'algorithme basé «utilisateur» [Samia, 2007].Il utilise toute la base de données[7] des évaluations des utilisateurs voisins de l'utilisateur actif « u » (les utilisateurs de la même communauté que u) ,pour calculer la prédiction que l'item « j » et pertinent ou non pour « u», on utilise l'approche centrée utilisateur, qui consiste à analyser la matrice "Utilisateur x Item" pour identifier des relations entre les utilisateurs (l'utilisateur actif « u » et les autres utilisateurs) et utiliser ces relations afin de calculer les prédictions. L'évaluation prédite est calculée par la fonction :

$$p_{a,j} = \overline{v_a} + k \sum_{i=1}^{n} w(a,i)(v_{i,j} - \overline{v_i}) \quad Avec \quad \overline{v_i} = \frac{1}{|I_i|} \sum_{j \in I_i} v_{i,j}$$

Où

$\overline{v_i}$ **:** L'évaluation moyenne des items pour l'utilisateur i

I_i : L'ensemble des items évalué par l'utilisateur i

n : Le nombre des utilisateurs voisins de « a ».

K : 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) [Ilham ,2010].

W (a, i) : Le poids calculé entre l'utilisateur actif « a » et les utilisateurs voisins « i » les détails de calcul des poids donnent lieu à des algorithmes différents, tel que l'algorithme basé sur la corrélation [Resnick, 1994] qui calcule le poids comme la corrélation entre les utilisateurs:

$$w(a,i) = \frac{\sum_j (v_{a,j} - \overline{v_a})(v_{i,j} - \overline{v_i})}{\sqrt{\sum_j (v_{a,j} - \overline{v_a})^2 \sum_j (v_{i,j} - \overline{v_i})^2}}$$

Et l'algorithme basé sur la similarité de vecteurs dans lequel le poids est calculé comme un cosinus entre les vecteurs [Brees, 1998] formés par les évaluations, des utilisateurs :

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

Une fois les prédictions calculées, le système de FC recommande à l'utilisateur actif « a » les items ayant les valeurs de prédiction les plus élevées.

[7] Sur tous les autres items

Cet algorithme présente les avantages suivants [Amokrane, 2007]:

✓ Simples à mettre en œuvre

✓ Les prédictions sont de bonne qualité[8]

✓ Peuvent être appliquées à tout type de document.

4.2.2.2. Algorithmes basés « modèle » :

Les méthodes basées sur un modèle ont été intégrées aux systèmes de recommandation pour remédier aux problèmes des méthodes basées sur la mémoire, dont notamment : la non-robustesse au manque de données ainsi que le non-passage à l'échelle [Sarwar et al., 2000b] [Su et Khoshgoftaar, 2009] .Elles utilisent notamment des techniques de réduction de dimensionnalité ou clustering dans le but d'écarter les utilisateurs ou les items non représentatifs et d'élaborer des modèles (généralement en hors ligne "off-line") qui serviront pour estimer la fonction de prédiction suivante :

$$p_{a,j} = E(v_{a,j}) = \sum_{i=0}^{m} \Pr(v_{a,j} = i \,|\, v_{a,k}, k \in I_a) i$$

Où m : les évaluations sur les items se fassent sur une échelle d'entiers de 0 à m

Le processus de construction du modèle est basé sur les techniques d'apprentissage automatique, telles que : le clustering, les réseaux bayésiens, etc.

a)*Le modèle de clustering :*

Un cluster est une collection d'objets qui sont similaires entre eux et dissimilaires aux objets appartenant aux autres clusters [Han et Kamber, 2001]. Les méthodes de Clustering permettent de limiter le nombre d'individus considérés dans le calcul de la prédiction d'un utilisateur sur un item, et de prédire les notes manquantes, par l'idée de regrouper en clusters homogènes les utilisateurs ayant les mêmes goûts, ou de regrouper en clusters les ressources portant sur les mêmes sujets, ou qui ont tendance à plaire aux mêmes personnes, pour arriver à une partition de base aussi pertinente que possible. Ensuite afin de calculer la prédiction on considère seulement le cluster de l'utilisateur courant, le temps de traitement sera donc plus court et les résultats seront potentiellement plus pertinents.

Les méthodes de clustering se différencient par deux stratégies de base la stratégie de contrôle et la fonction objective:

[8] Car sont recalculées à chaque fois pour chaque utilisateur dans la communauté

✓***Les stratégies de contrôle*** **:** Elles sont utilisées pour lancer une optimisation itérative du modèle courant afin de parcourir l'espace des clusters possibles. Cette stratégie est basée sur deux méthodes stratégie basé-utilisateur et celle basé-item :

a) Les stratégies de contrôle basé utilisateurs « les Clustering d'utilisateurs» : Elles consistent à créer des clusters de taille fixe ou variable (figure 2.10), qui regroupent les utilisateurs similaires par rapport à leurs profils .Plusieurs algorithmes sont construits pour le faire qui sont utilisés pour calculer la prédiction *en ligne* :

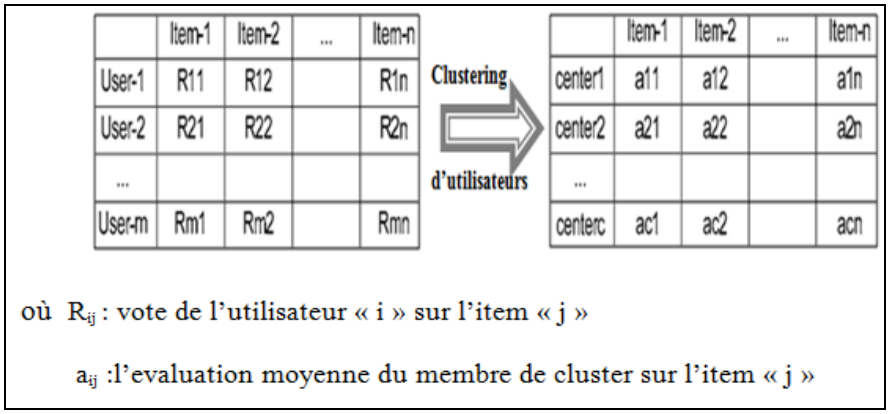


Figure 2.10 : Principe de la méthode basé utilisateurs

Parmi ces algorithmes, nous citons :

✓**La méthode des plus proches voisins « K-Means »** [Herlocker, 1999]:Est la méthode la plus courante, présenter dans la figure 2.11 :

Algorithm 1 Algorithme de partitionnement k-means

1: **Input** : k : le nombre de clusters et M : matrice "Utilisateur x Item"
2: **Output** : k clusters

3: Choisir aléatoirement k centroïdes initiaux de clusters
4: **repeat**
5: Réaffecter chaque utilisateur au cluster auquel il est le plus similaire
6: Recalculer les distances des utilisateurs dans chaque cluster
7: Mettre à jour les centroïdes
8: **until** Stabilité des centroïdes

Figure 2.11 :L'algorithme de partitionnement K-Means

Elle consiste dans un premier temps à choisir aléatoirement k centres « k utilisateurs » dans l'espace de représentation comme des utilisateurs actifs. Ensuite, chaque utilisateur de l'espace de représentation est mis dans le cluster du centre le plus proche tel que la distance entre cet utilisateur et le centre du cluster est faible .La métrique utilisée pour déterminer la distance par rapport à ces centres est le coefficient de corrélation de Pearson.

Après, en prenant en compte les utilisateurs qui viennent d'être affectés aux clusters, les distances des utilisateurs dans chaque cluster sont recalculées afin de trouver la position du centre de chaque cluster c.à.d. découvrir un autre centre.

Après la découverte des nouveaux centres, les distances sont à nouveau réévaluées afin de retrouver le cluster auquel chaque utilisateur devrait appartenir. Cette opération est itérée jusqu'à ce que les centres deviennent stables et ne changent plus.

Ensuite, à chaque arrivée d'un nouvel utilisateur actif, le système génère des prédictions «on - line» en cherchant dans les clusters, cela qui contient l'utilisateur courant par l'application du coefficient de corrélation de Pearson entre l'utilisateur central de chaque cluster et l'utilisateur actif , et en appliquant ensuite la formule définie dans la section «algorithme basé- mémoire» où l'ensemble des utilisateurs considéré se limite à ceux de cluster.

Parmi ses avantages :

- ✓ Efficacité [Su et Khoshgoftaar, 2009]
- ✓ Simples à mettre en œuvre
- ✓ Évoluer dynamiquement en fonction des profils utilisateurs

✓**PAM « Partitioning Around Medoïds[9] » :**

C'est une méthode de clustering, proposée afin de réduire la sensibilité aux données aberrantes et de remédier au problème du recouvrement des clusters [Han et Kamber, 2001]. Le principe de cette méthode est de créer un ensemble de clusters, tels que dans chaque cluster, tous les utilisateurs ont la même priorité pour représenter un utilisateur central appelé "médoïde". La figure 2.12 présente l'algorithme qui décrit les étapes du clustering PAM [Han et Kamber, 2001].

L'algorithme PAM itère jusqu'à ce que les centres deviennent stables, i.e., jusqu'à ce que les u_{med} ne changent pas. Durant cette itération, la qualité du clustering est évaluée en utilisant une fonction qui calcule le coût total S. Ce coût mesure l'erreur en cas de permutation d'un médoïde initial u_{med} avec un autre médoïde u_{random}. Si S est négative, u_{med} est remplacé effectivement par u_{random}. Autrement, u_{med} est considéré comme acceptable et devient stable.

[9] Un utilisateur central

Algorithm 2 Algorithme de partitionnement PAM

Input : k : le nombre de clusters et M : matrice "Utilisateur x Item"
Output : k clusters

Choisir aléatoirement k utilisateurs comme étant les médoïdes initiaux de clusters
repeat
    Affecter chaque utilisateur à un cluster tel que la dissimilarité entre cet utilisateur et le médoïde est faible
    Sélectionner aléatoirement un utilisateur non-représentatif (non-médoïde) u_{random}
    Calculer le coût total, S, de permutation d'un utilisateur représentatif u_{med} avec u_{random}
    if $S < 0$ **then**
        Remplacer u_{med} par u_{random} pour former les nouveaux médoïdes
    end if
until Stabilité des médoïdes

Figure 2.12 : Algorithme de partitionnement PAM

Son principe de fonctionnement est semblable au précédent, à l'arrivée d'un nouvel utilisateur actif. Son principe avantage est l'insensibilisation aux données aberrantes : un centre constitue l'objet ou l'utilisateur le plus central du cluster. Ceci est assuré en permutant systématiquement un centre et un autre utilisateur choisit aléatoirement afin de vérifier si la qualité du clustering décroît [Tufféry, 2007].

b)Les stratégies de contrôle basé items « les clustring d'items »: Quand le système manque de données, la modélisation des utilisateurs devient difficile et complexe. En effet, dans le cadre du filtrage collaboratif, le système serait incapable d'identifier un nombre significatif de voisins nécessaires au calcul de recommandations adaptées aux besoins de l'utilisateur actif, pour cela le FC utilise un autre principe de clustring basé items. Ils consistent à créer des clusters de taille fixe ou variable, qui regroupent des items qui semblent d'avoir des évaluations semblables. Une fois que les clusters sont créés, des prévisions pour un item peuvent être faites, en cherchant dans les clusters, cela qui contient l'item courant et en faisant la moyenne des avis des autres items. Les systèmes [Amokrane, 2007] ou le nombre d'utilisateurs est plus important que le nombre de ressources, donc il est préférable de baser sur la classification des ressources pour améliorer la qualité de la prédiction. Les expérimentations de [Sarwar ,2001] ont montré que ce type d'approche pouvait améliorer les performances de 28 fois et la qualité de 27%.Plusieurs algorithmes sont construit afin de créer des clusters, parmi ces algorithmes, nous citons K-means et l'algorithme de Greg et al :

✓**K-means :**

Elle consiste dans un premier temps à choisir aléatoirement k centres « k items » dans l'espace de représentation .En suite, chaque item de l'espace de représentation est mis dans le

cluster du centre le plus proche tel que la distance entre cet item et le centre du cluster est faible .La métrique utilisée pour déterminer la distance par rapport à ces centres est le coefficient de corrélation de Pearson. Après, en prenant en compte les items qui viennent d'être affectés aux clusters, les distances des utilisateurs dans chaque cluster sont recalculées afin de trouver la position du centre de chaque cluster c.à.d. découvrir un autre centre .Après la découverte des nouveaux centres, les distances sont à nouveau réévaluées afin de retrouver le cluster auquel chaque item devrait appartenir. Cette opération est itérée jusqu'à ce que les centres deviennent stables et ne changent plus. En suite, à chaque fois quant a un item actif, le système génère des prédictions , en cherchant dans les clusters, cela qui contient l'item courant par l'application de le coefficient de corrélation de Pearson entre l'item central de chaque cluster et l'item actif et en appliquant ensuite la formule suivante de Weighted Sum :

$$P_{ut} = \frac{\sum_{i=1}^{c} R_{ui} \times sim(t,i)}{\sum_{i=1}^{c} sim(t,i)}$$

Où $R_{u,i}$:l'evaluation de l'utilisateur « u » sur l'item « i »

✓Greg et al [Linden, 2003] sont proposés un algorithme de clustering .Il est utilisé dans le domaine de commerce électronique « Amazon 2003 » afin de résoudre le problème de passage a l'échelle et d'améliorer les performances de système de filtrage utiliser .Le principe de ce algorithme est de créer des clusters qui regroupent les items qui sont achetés en ensemble par les utilisateurs, la fonction utilisée pour calculer la similarité entre les items est de Cosinus :

$$sim(i,j) = \cos(\vec{i},\vec{j}) = \frac{\vec{i} \cdot \vec{j}}{||\vec{i}||_2 * ||\vec{j}||_2}$$

Où $\vec{i}$: le vecteur d'évaluations de l'item actif i

✓***Les fonctions objectives :*** L'idée du clustering est alors de former des clusters d'utilisateurs et/ou des clusters d'items les plus pertinents et significatifs possibles .Pour estimer la qualité de ces clusters, des fonctions objectif sont utilisées qui permettre d'évaluer la pertinence de la partition effectuée. Parmi ces fonctions :

•L'idée ici est de favoriser les clusters qui vont accroître la prédictibilité, c'est à dire

$$P(V_{ij}=k \mid i \in C_n) > P(V_{ij}=k)$$

Où V_{ij} : La note attribuée par les utilisateurs (i) sur l'article (j),

C_n : Le cluster « n ».

•Une autre idée est de minimiser la distance moyenne entre éléments d'un même cluster tout en maximisant la distance entre clusters, pour déterminer la qualité d'un cluster trois étapes sont suivis :

✓On définit d'abord le centre X_0 du cluster par:

$$\overline{X_o}=\frac{\sum_{i=1}^{N}\overline{v_i}}{N}$$

✓Puis deux mesures sont proposées ici pour définir la distance entre éléments d'un même cluster, ces deux mesures doivent alors être minimisées pour maximiser la similarité des observations à l'intérieur d'un cluster :

•Le radius R correspond à la distance moyenne entre les membres du cluster et son centre:

$$R=\left(\frac{\sum_{i=1}^{N}(\overline{v_i}-\overline{X_o})^2}{N}\right)^{1/2}$$

•Le diamètre *D* correspond à la distance moyenne entre paires de membres du cluster:

$$D=\left(\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(\overline{v_i}-\overline{v_j})^2}{N(N-1)}\right)^{1/2}$$

✓Ensuite, deux méthodes sont proposées ici pour définir la distance entre deux clusters et mesurer leur proximité, ces deux mesures doivent alors être maximisées pour minimiser la similarité des observations entre clusters:

•Etant donné les centres X_{0i} de deux clusters, la distance euclidienne D_0 entre centres est donnée par:

$$D_0=\left((\overline{X_{0_1}}-\overline{X_{0_2}})^2\right)^{1/2}$$

•Et D_1, la distance de Manhattan entre centres est donnée par:

$$D_1=|\overline{X_{0_1}}-\overline{X_{0_2}}|=\sum_{i=1}^{d}|\overline{X_{0_1}}^{(i)}-\overline{X_{0_2}}^{(i)}|,$$

b) Le modèle réseaux bayesiens [Mooney, 1998]: Dans le modèle à base de réseaux bayesiens, les nœuds sont correspondent aux documents, les états de chaque nœud correspondent aux valeurs d'évaluation possibles. Il inclut également un état correspondant à l'absence d'évaluation pour les domaines où il n'y a pas d'interprétation naturelle des

données manquantes. Le réseau est ainsi construit à partir des données en appliquant un algorithme d'apprentissage, le résultat est alors sous forme d'arbres de décision représentant chaque table de probabilité conditionnelle pour chaque nœud [Houda, 2008]. 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 données aux articles parents (le même principe de cluster) présents dans le réseau, et on attribue, pour l'article considéré, la note la plus probable [Amokrane, 2007].Parmi ses avantages [Houda, 2008]**:**

- ✓Traiter le problème du manque d'évaluations, ceci en regroupant les utilisateurs et les documents en groupes ou classes
- ✓Peut être appliqué à tout type de document

4.2.2.3. Algorithmes basés sur un «apprentissage automatique»:

Cet algorithme propose l'utilisation d'un agent de prédiction pour chaque utilisateur, chaque agent doit opérer sur un ensemble d'essai avec une prédiction à faire à chaque étape. Les algorithmes d'apprentissage en ligne « reposent sur le principe d'apprentissage à partir d'avis d'experts » .Ils sont un processus continu et interactif et reposent sur un ensemble d'algorithmes, considérés comme des «experts prédicateurs» qui se présentent sous forme d'un ensemble d'agents de prédiction indépendants. Ces experts sont associés à des poids qui mesurent leur confiance envers la tâche de prédiction qu'ils réalisent. Chaque agent sera confronté à un ensemble d'essais avec une prédiction à faire à chaque étape en fonction des agents qui l'entourent et qui pourront avoir un comportement similaire , neutre ou opposé à la fonction que cet agent cherche à atteindre[Samia, 2007] .Parmi ses points forts :

- ✓Peuvent être appliqués à tout type de document
- ✓Les prédictions sont de bonne qualité.

4.2.3. Les problématiques de la production des recommandations :

La production des recommandations connait plusieurs problèmes parmi en :

- ✓ *Le démarrage à froid pour un nouveau document* qui ne contient aucune évaluation
- ✓ *La complexité du traitement de l'algorithme basé-mémoire* et se complexifie avec l'augmentation du nombre d'utilisateurs et de ressources [Houda, 2008]
- ✓ Le faible nombre de ressources communément évaluées par les utilisateurs engendre des prédictions peu pertinentes « problème de matrice creuse » [Houda, 2008]:
- ✓ Les algorithmes basé-modèle deviennent non pratiques pour une large base de données [Houda, 2008].

- ✓ *Matrice creuse:* Des solutions sont proposées pour résoudre ce problème où, on peut exploiter le technique basé sur le contenu [Lang,1995] [Krulwich et Burkey, 1996] [Billsus et Pazzani, 2000], lorsqu'un nouvel item est introduit, le système évalue la similarité de contenu de cet item avec les items disponibles afin de l'impliquer au processus de recommandation. Ou l'utilisation du technique à base d'agents, [Park, 2006] propose l'utilisation des agents comme moyenne d'évaluations automatiques
- ✓ *l'effet entonnoir pour les algorithmes basé-modèle :* La majorité des systèmes actuels ne permettent pas dans certains cas de prendre en compte les documents d'un nouvel axe de recherche pour les différentes communautés

4.3. L'évaluation des recommandations :

L'évaluation est le jugement porté sur un document, aide le système pour former les communautés. Elle peut être capturée de différentes manières :

✓***Evaluation immédiate :*** Document évalué juste après l'examination de son contenu

✓***Evaluations à posteriori [Amokrane, 2007] :*** Il faut prendre en considération la durée d'utilisation de la ressource nécessaire pour pouvoir recueillir une évaluation.

✓***Evaluation implicite [An ,2006] :*** Le système induit la satisfaction de l'utilisateur à travers ses actions. Par exemple, le système estimera qu'une recommandation supprimée correspond à une évaluation très mauvaise, alors qu'une recommandation imprimée ou sauvegardée peut être interprétée comme une bonne évaluation.

✓***Explicite [An, 2006] :***L'utilisateur donne une valeur numérique sur une échelle donnée ou une valeur qualitative de satisfaction, par exemple, mauvaise, moyenne, bonne et excellente.

5. Exemples des systèmes de filtrage collaboratif:

Avec l'avènement de l'Internet et des applications web, il y a eu un engouement pour les systèmes de recommandation et sur tous les systèmes de filtrage collaboratif qui se sont développés dans différents domaines d'application. Nous pouvons en citer :

✓*Tapestry :*

Le concept du filtrage collaboratif a été lancé avec le projet Tapestry à Xerox Parc. La gestion des e-mails est sa motivation première [Goldberg, 1992]. Tapestry repose sur une «recommandation commentée» basé sur des annotations de qualité ou d'appréciation des documents faites par les utilisateurs. De cette manière, les documents sont filtrés en fonction de ces annotations [Lumineau, 2002].

✓ *GroupLens :*

GroupLens [Resnick, 1994] ; [Miller, 1997], est un système expérimental de l'université du Minnesota, il est un des plus célèbres et solides dans ce domaine. Il est semblable dans son esprit à Tapestry : les lecteurs sont appelés à noter les articles qu'ils lisent. Le système trouve alors des corrélations entre les différents utilisateurs et identifie des groupes d'utilisateurs dont les intérêts sont semblables. Ensuite, il emploie ces estimations pour prédire l'intérêt que porteront les utilisateurs à chaque article.

✓ *Le système de Malltz et Ehrlich :*

Le système de Maltz et Ehrlich est présenté comme un substitut au mail dans ces situations. Il est intégré à un système de recherche d'informations et permet à ses utilisateurs d'adresser des pointeurs aux personnes qu'ils jugent intéressées, sans avoir à interrompre leur session de recherche d'informations. D'un autre côté, l'ensemble de ces échanges est stocké pour constituer une base de références [Berrut, 2003].

✓*Amazon:*

Amazon utilise l'algorithme « item-to-item collaborative filtering » [Linden, 2003]. Ce système commence par le calcul du degré de similarité entre articles en hors ligne « offline » construisant ainsi une table des similarités articles. Cette étape est extrêmement gourmande en termes de temps de calcul. Ensuite, si l'utilisateur s'intéresse à un produit bien précis, le système lui recommande des produits similaires à celui-ci sur la base de la matrice des similarités des articles.

6. Conclusion :

Le filtrage collaboratif consiste à filtrer les documents du flux entrant en se basant sur les profils de chaque utilisateur. En générale le processus de filtrage collaboratif et basé sur l'un des deux algorithmes basé modèle ou basé mémoire où ce dernier utilise toute la base de données des évaluations des utilisateurs voisins de l'utilisateur actif « u » ,pour calculer la prédiction que un item « j » et pertinent ou non pour un utilisateur « u», par contre Les méthodes basées modèle ont été intégrées aux systèmes de recommandation pour remédier aux problèmes des méthodes basées sur la mémoire, dont notamment : la non-robustesse au manque de données ainsi que le non-passage à l'échelle. Ils utilisent notamment des techniques de réduction de dimensionnalité ou clustering dans le but d'écarter les utilisateurs ou les items non représentatifs et d'élaborer des modèles qui serviront pour estimer la fonction de prédiction, le processus des systèmes de basé modèle sont exécutés on trois

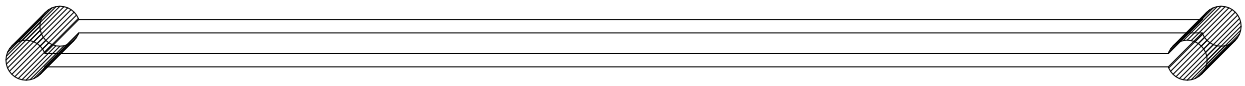
étapes : *la formation des communautés*, la *production des recommandations*, et l'évaluation *des recommandations*. Le tableau 2.1 présente une fiche comparative du technique de recommandation basé mémoire et basé modèle :

Catégorie	Des exemples	Avantages	Inconvénients
FC[10]. basé mémoire	FC exploitant : •l'approche k-Means (basé utilisateurs ou basé items) •des mesure de Pearson	✓Implémentation simple ✓Intégration simple de nouvelles données ✓Précision de recommandations	✖ Dépendance aux données de notes ✖ Détérioration de la qualité de recommandation à cause du manque de données ✖ Problème de passage à l'échelle
FC. basé modèle	FC exploitant : •Clustring •Aproche probabiliste	✓Remédier aux problèmes des méthodes basées sur la mémoire ✓Amélioration de la qualité des recommandations ✓Réduction de problème de manque de données ✓Prédiction des futures comportements de navigation	✖ Construction coûteuse de modèles ✖ Risque de perte d'informations dû à la réduction de dimensionnalité ✖ Problème de l'effet entonnoir

Tableau 2.1 : synthèse comparative des techniques de recommandation

[10] Filtrage collaboratif

Chapitre 3



Le web sémantique et le filtrage collaboratif

1. Introduction :

La majorité du contenu du web produit est conçu pour être lu par des êtres humains, et pas pour être manipulé symboliquement par des programmes informatiques. Certes un document HTML est manipulé par un programme pour que la mise en page soit correcte. Mais ce traitement se limite à interpréter les balises de présentation HTML présentes dans le document. Ces balises se limitent à décrire la manière dont le document doit être présenté. La signification du contenu du document reste implicite et le document ne peut donc pas être manipulé sur base de cette signification. De manière générale les ordinateurs n'ont aucune méthode systématique pour traiter le contenu d'un document web sur la base de leur sémantique.

Dès 1994 une nouvelle génération de web sémantique est développée par Tim Berners-Lee. A pour ambition de lever cette difficulté. Les ressources du web seront plus aisément accessibles aussi bien par l'homme que par la machine, grâce à la représentation sémantique de leurs contenus. Le web sémantique est intégré dans plusieurs systèmes comme une ressource fournissant des informations supplémentaires, permettant d'enrichir ces systèmes et d'améliorer leurs performances.

Dans ce chapitre nous définissons tout d'abord la notion de web sémantique, ensuite nous présentons les travaux qui intègrent le web sémantique dans le processus de filtrage collaboratif : la formation des communautés, La production de recommandations et l'évaluation de recommandation. Enfin, nous terminons par une conclusion sur l'influence de cette nouvelle notion sur la performance des systèmes de filtrage collaboratif.

2. Web sémantique :

2.1. Présentation :

L'expression web sémantique, due à Tim Berners-Lee au sein du W3C, fait d'abord référence à la vision du web de demain comme un vaste espace d'échange de ressources entre êtres humains et machines permettant une exploitation, qualitativement supérieure, de grands volumes d'informations et de services variés. Il devrait voir, à la différence du web que nous connaissons aujourd'hui, les utilisateurs déchargés d'une bonne partie de leurs tâches de recherche, de construction et de combinaison des résultats, grâce aux capacités accrues des machines à accéder aux contenus des ressources et à effectuer des raisonnements sur ceux-ci [Laublet,2004].

Parmi les outils de web sémantique l'ontologie qui fournit la base sémantique du web sémantique, elle [Gruber, 1993] définit le vocabulaire partagé pour aboutir à une compréhension commune d'un domaine donné (la figure 3.1 présente un exemple d'une ontologie de films). De nombreux langages sont définis pour représenter les ontologies comme OWL (Web Ontology Language), « protégée »… etc, ils sont utilisés pour représenter la signification des termes d'un vocabulaire et les relations entre ces termes.

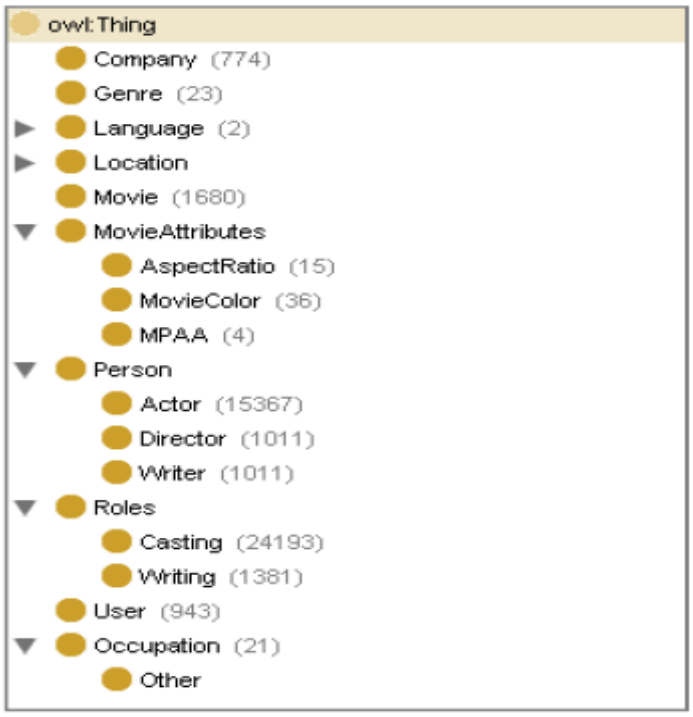


Figure 3.1 : Exemple d'une ontologie des films [Amokrane, 2007] [Latifa, 2010]

Les annotations décrivent le contenu de la ressource et en plus fournissent des informations contextuelles et sémantiques. Les systèmes d'annotations offrent à leurs utilisateurs la possibilité d'héberger leurs photos, vidéos, documents ou toute autre ressource et surtout leur assigner un ensemble de mots décrivant leurs contenu : c'est ce qu'on appelle les annotations ou tags [Houda,2008].

Sans oublier le réseau social qui représente une structure sociale dynamique, se modélisant par des sommets et des arêtes. Les sommets désignent généralement des gens et/ou des organisations et sont reliés entre eux par des interactions sociales. Il aide à créer un cercle d'amis, à trouver des partenaires commerciaux, un emploi ou autres. Il s'agit de services de réseautage social, comme Facebook, Twitter, Identi.ca…etc.

Les annotations sont souvent associées au web sémantique du fait qu'elles permettent aux utilisateurs d'ajouter une métadonnée aux documents d'une manière très simple [Houda, 2008].

2.2. Vision :

La vision du web sémantique est une vision globale, large. Les visions peuvent être perçues sous différents angles, selon l'interaction faite avec l'infrastructure [Passin, 2004].

✓Rendre les données compréhensibles par des machines : La vision est d'avoir des ressources web pouvant être exploitées par les machines, non pour l'affichage seulement.

✓Pouvoir utiliser des agents intelligents : Cette vision étend la première en y ajoutant la notion d'agents intelligents, ceux-ci pourront avoir des raisonnements logiques sur les ressources.

✓Vision de base de données partagée : C'est une vision très intéressante car elle étend la vision du web classique. Le web sémantique vise à reproduire sur les connaissances ce que le web à fait pour les données (documents destinés à la lecture).

✓Vision d'infrastructure automatisée : Le web tel quel ne permet pas d'automatisé. Les seules choses que l'on peut faire sont : gérer des liens, indexer des pages … Le web sémantique n'étant pas une application mais une infrastructure, vise à fournir des briques qui permettront de faire plus de choses automatiquement.

✓Annotations améliorées : Avec le web sémantique, il sera possible d'avoir des annotations exprimées dans un langage compréhensible par les machines.

✓Améliorer la recherche d'informations : Permet d'avoir des recherches plus sophistiquées pas seulement par des mots clés mais par la navigation et le contenu.

✓Services web : Par cette vision, le web sémantique permettra de partager des services et de pouvoir rechercher ces services.

3. Le filtrage collaboratif et le web sémantique :

L'introduction des aspects du web sémantique dans les systèmes de filtrage « collaboratif » a été au centre des discussions ces dernières années produisant une nouvelle génération de systèmes de filtrage enrichis par la sémantique ou les aspects sociaux du web sémantique, c'est ce qu'on appelle les systèmes de filtrage collaboratif sémantique ou social.

Les techniques du web 2.0 peuvent intégrer dans n'importe quelles étapes dans le processus de fonctionnement des systèmes de filtrage collaboratif, la formation des communautés, La production de recommandations et l'évaluation de recommandation :

3.1. La formation des communautés :

Consiste à regrouper les utilisateurs qui ont des propriétés « profils » communes, dans une même communauté, ce processus est réalisé en deux étapes : le regroupement des profils des utilisateurs et la formation des communautés:

3.1.1. Le regroupement des profils :

Un profil utilisateur est une source de données qui contient un ensemble d'informations concernant divers aspects de l'utilisateur pouvant être utilisées pour adapter et/ou dicter le comportement du système [Amokrane ,2007].Les profils sont capturés de deux manières explicite représentés par les acquis de l'utilisateur, ses connaissances, ses objectifs, ses préférences ou implicite induites à partir de l'activité de navigation, d'évaluation, réponse à des questions, etc. Avec l'apparition du web 2.0 et ses technique d'ontologie, d'annotation et de réseau social, des travaux ont montré l'importance d'intégrer ces aspects sémantiques qui permettront d'apporter de l'information supplémentaire pour réduire l'effet du démarrage à froid[Amokrane ,2007],et d'améliorer la performance des systèmes de filtrage d'informations. Parmi ces systèmes de filtrage d'information on a :

✓ *le système Entrée* (qui recommande des restaurants) [Burke, 2002] : utilise une méthode de combinaison du filtrage collaboratif avec le filtrage basé sur le contenu, afin de résoudre le problème de démarrage à froid .Le système demande à l'utilisateur de définir ses centres d'intérêt, en termes de contenu, à partir d'une liste de termes et/ou d'exemples décrivant au mieux ses centres d'intérêt.

✓ Middleton et Shadbolt [Middelton, 2002] : Dans leur système de recommandation d'articles de recherche, ont exploité les ontologies pour identifier les groupes d'utilisateurs ayant des caractéristiques en commun. C'est ce qu'ils ont appelé communautés en pratique (Communities of Practice CoP) [Middelton, 2003].

✓ *le système Quickstep-Foxtrot* (qui recommande des papiers scientifiques) [Middleton et al., 2004] utilise une technique de classification couplée avec une représentation ontologique des domaines de recherche afin d'extraire les centres d'intérêts de l'utilisateur .Il utilise une ontologie fournit l'historique *des publications des utilisateurs* ,ainsi qu'une liste classée des utilisateurs similaires. Il assigne chaque article à une classe (thème) avec laquelle le vecteur représentatif du document est le plus similaire. Le profil initial est construit à partir de la corrélation de ses publications et les profils similaires. Cette approche se base essentiellement sur l'ensemble des publications de l'utilisateur pour construire son profil, ceci désavantage les utilisateurs n'ayant pas encore publié d'articles ainsi que ceux dont les centres d'intérêts actuels (projets en cours) diffèrent de ceux collectés de leurs anciennes publications.

✓ Houda [Houda,2007] propose une approche hybride de construction de profil des utilisateurs .Le processus de construction est composé de trois dimensions :

- *Dimension collaborative :* contient l'ensemble des évaluations données par l'utilisateur, elles sont recueillies au fur et à mesure que l'utilisateur évalue des documents soit de manière explicite ou implicite .L'ensemble des vecteurs d'évaluations de tous les utilisateurs constitue la traditionnelle matrice Utilisateurs/Documents.

- *Dimension sociale :* contient l'ensemble des données personnelles relatives à l'utilisateur : nom, prénom, date de naissance, sexe, profession, Email, page web personnel, l'ensemble des contacts d'ordre professionnel et /ou personnel (liste d'amis), ces derniers peuvent être directement donnés par l'utilisateur ou calculés avec la formule de corrélation de Pearson [Breese, 1998]

$$w(u,u_i)=\frac{\sum_j (v_{u,j}-\bar{v}u)(v_{i,j}-\bar{v}i)}{\sqrt{\sum_j (v_{u,j}-\bar{v}u)^2 \sum_j (v_{i,j}-\bar{v}i)^2}}$$

Avec :

u : L'utilisateur en cours

u_i : Le contact de l'utilisateur u

j : L'ensemble des documents sélectionnés par le module *Social Matching* et évalués à la fois par u et u_i

Les données sont collectées dès leur inscription et peuvent être mises à jour au cours du temps;

- *Dimension sémantique :* Traduit les centres d'intérêts de l'utilisateur. Elle est représentée sous forme de concepts ou thèmes avec des poids reflétant leurs degrés d'importance vis-à-vis de l'utilisateur. Ces concepts sont tirés d'une hiérarchie générale de concepts ou de domaine souvent appelés profil général. Le poids du concept est donné par la formule suivante :

$$v(c)=\frac{\sum_j w_j v_{u,j}}{\sum_j w_j}$$

Avec : w_j : Le poits su concept « c » dans le document « j »

$V_{u,j}$: La note donnée au document j

Cette fonction est normalisée pour avoir des valeurs entre [0,1] par :

$$w(c)=\frac{v(c)}{Max(v)}$$

3.1.2. Formation des communautés :

Consiste à regrouper les utilisateurs similaires dans les même communautés (qui on des propriétés communes) par le calcul de la similarité. Le web sémantique avec ses notions (ontologie, annotation…) fournit des informations supplémentaires permettant d'enrichir les similarités collaboratives calculées entre les utilisateurs qui deviennent une similarité[11] sémantique.les approches proposées afin de la calculer sont classées en quatre catégories:

✓ ***Comptage des arcs entre deux termes (Edge Counting) :*** La méthode la plus intuitive pour le calcul de la similarité dans une ontologie est le calcul du nombre d'arcs entre deux concepts dans la hiérarchie. Cette Méthode est utilisée lorsque l'ontologie est sous forme d'arborescence où la relation « est un » est utilisée, La similarité entre deux termes est calculée en fonction de la distance qui lie les deux termes dans la hiérarchie (plus proche parent commun). La distance la plus intuitive est la distance proposée par Rada [Rada ,1989]. La distance entre deux concepts est représentée par la distance minimale entre les éléments et le parent commun le plus récent.

$$\mathrm{sim}(c_1, c_2) = \frac{1}{1 + dist(c_1, c_2)}$$

L'exemple suivant démontre comment on calcule dist (c_1, c_2) :

La distance entre 'Professor' et 'Administrator' est égale à la distance entre 'Professor' et 'Employee' plus la distance entre 'Administrator' et 'Employee', Le concept 'Employee' étant le parent commun le plus récent.
La distance est ainsi égale à 3. La distance entre 'Professor' et 'Student' est égale à 4.

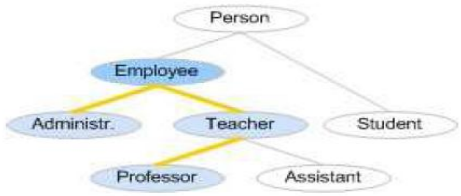


Figure 3.2 : Exemple d'un comptage des arcs entre deux termes

Une première possibilité pour la représentation de la similarité est mentionnée par Resnik [Resnik ,1999]. La similarité de Resnik est donnée par la formule suivante :

$$\mathrm{sim}_{\mathrm{edge}}(c_o^x, c_o^y) = \frac{2 * MAX - len(c_o^x, c_o^y)}{2 * MAX}$$

Max : Représente la plus grande distance entre la racine de l'ontologie et les feuilles de l'arborescence.

$len(c_o^x, c_o^y)$: La plus petite distance entre deux concepts Cx et Cy

[11] On aura toujours une valeur dans [0, 1]

✓ ***Approche basée sur le contenu informationnel (Information Content)*** : Le problème de la distance dans une ontologie est qu'elle dépend du caractère subjectif de la construction des ontologies. Pour traiter ce problème, Resnik [Resnik,1999] défini la similarité entre deux concepts (dans le cas des termes) comme étant :

$$\mathrm{sim_{res}}(c_o^x, c_o^y) = max_{c_o^z \in c_A(c_o^x, c_o^y)}[-\log P_{c_o^z}]$$

$CA(Cx, Cy)$: Est l'ensemble des ancêtres communs de Cx et Cy

$-\log P_{c_o^z}$:Le contenu en informatique du terme z

Où PC_z est la probabilité d'apparition du terme z dans le corpus de référence.

Lin [Lin 98] défini la similarité différemment :

$$\mathrm{sim_{lin}}(c_o^x, c_o^y) = \frac{2 * logP_{M\,RCA(c_o^x, c_o^x)}}{log\ P_{c_o^x} + logP_{c_o^x}}$$

$logP_{M\,RCA(c_o^x, c_o^x)}$: Le contenu en informatique du parent commun le plus récent

$log\ P_{c_o^x}$:Le contenu en informatique du terme x

✓ ***Similarité vectorielle :*** Les termes d'une ontologie peuvent être représentés par des vecteurs de valeurs, de mots, d'évaluations, etc. Plusieurs métriques[Amokrane, 2007] de similarité peuvent être utilisées pour calculer les distances entre les vecteurs. Parmi les distances typiques on trouve :

Cosinus

$$sim_{cosine}(\vec{x}, \vec{y}) = \frac{\vec{x} \bullet \vec{y}}{|\vec{x}| \times |\vec{y}|}$$

Jaccard

$$sim_{jaccard}(\vec{x}, \vec{y}) = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} x_i y_i}$$

Overlap

$$sim_{overlap}(\vec{x}, \vec{y}) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}}$$

Dice

$$sim_{dice}(\vec{x}, \vec{y}) = \frac{2 * \sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2}$$

Euclidienne

$$d_{euclid}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

$$sim_{dist}(\vec{x}, \vec{y}) = \frac{1}{1 + d_{dist}(\vec{x}, \vec{y})}$$

✓ ***Méthodes hybrides :*** Dans les approches hybrides on combine les approches définies précédemment et ceci pour améliorer la précision quand une approche unique est insuffisante.

[Mobasher, 2004] propose une mesure de similarité qui combine la similarité sémantique entre documents avec celle basée sur les votes :

$$CombinedSim(i_p,i_q)=\alpha \cdot SemSim(i_p,i_q)+(1-\alpha) \cdot RateSim(i_p,i_q)$$

Avec :

α Un paramètre de combinaison sémantique spécifiant le poids de la similarité sémantique dans la combinaison linéaire.

[Houda, 2007] Propose une représentation à trois dimensions du profil utilisateurs : dimension collaborative, dimension sociale, dimension sémantique. Chacune de ces dimensions est exploitée, pour inférer des communautés collaboratives, sociales et sémantiques. La communauté collaborative classique se base sur la similarité des votes, la communauté sociale contiendra l'ensemble des amis de l'utilisateur et les amis de ses amis et enfin, la communauté sémantique, peut être calculée à partir des scores de similarité sémantique globale en tenant compte de chaque centre d'intérêt à part.

3.2. La production de recommandations :

Ce processus est exécuté à l'arrivée d'une nouvelle information ou si un nouveau document « new item » ou un nouvel utilisateur est intégré dans une communauté « new user » :

3.2.1. La production des recommandations cas de nouveau document:

Généralement ce problème est traité en combinant une approche de filtrage basé sur le contenu avec le filtrage collaboratif « approche hybride », par exemple en utilisant la similarité, au niveau du contenu, entre documents pour estimer la satisfaction des utilisateurs sur le nouveau document en fonction de leurs évaluations sur certains documents assez proches c.à.d. voisins [Schein, 2001];ou en introduisant des agents intelligents qui évaluent les documents automatiquement [Good ,1999]. Middelton [Middelton, 2004] propose un algorithme pour les nouveaux documents, Les documents sont représentés sous forme de vecteurs où chaque valeur représente le poids d'un terme de l'article. Le poids étant égal à la fréquence du mot divisé par le total des mots de l'article, ainsi on peut avoir des vecteurs de dimension égale à 15000 après réduction. Ensuite, on définit son voisinage à l'aide d'un algorithme de classification: k-Nearest Neighbor [Aha, 1991] (figure 3.3) renforcé par celui de [Freund, 1996]: AdaBoostM1.

$$w(d_a, d_b) = \sqrt{\sum_{j=1..T} (t_{ja} - t_{jb})^2}$$

Avec :

$w(d_a, d_b)$	KNN distance entre les documents a et b
d_a, d_b	Les vecteurs des documents a et b
T	Nombre de termes dans l'ensemble des documents
t_{ja}	Poids du terme *j* dans le document *a*

Figure 3.3: L'algorithme du K-Nearest Neighbor

Amokrane [Amokrane, 2007] propose un algorithme qui se base spécialement sur les calculs en arrière plan qui réagissent bien au démarrage à froid. La solution d'Amokrane repose sur l'adoption d'un système hybride où le calcul des similarités se base sur deux aspects, l'aspect collaboratif et l'aspect sémantique. L'aspect sémantique permettra d'apporter de l'information supplémentaire pour réduire l'effet du démarrage à froid, nouvel utilisateur, nouvelle ressource et des matrices creuses.

L'algorithme d'Amokrane :

✓Un moteur collaboratif, pour le calcul des similarités, et l'ajout de nouvelles évaluations en recalculant un certain nombre de distances, lorsqu'un utilisateur 'U' évalue une ressource 'R', il faut recalculer toutes les distances entre l'utilisateur 'U' et les utilisateurs ayant évalués la ressource 'R' ainsi que les distances entre 'R' et toutes les ressources étant évaluées par 'U'.

✓ Un moteur de similarité sémantique, pour le calcul des similarités sémantique.

✓ Un module de clustering (éventuellement pour monter en charge).

✓ Un moteur de prédiction et de recommandation.

3.2.2. La production des recommandations cas de nouvel utilisateur:

On rappelle que pour la production des recommandations trois techniques sont utilisées : basés-mémoire, basés-modèle et basé sur un «apprentissage automatique».Deux grands problèmes peuvent arrivés dans l'application de ces algorithmes : la matrice creuse et le passage à l'échelle:

✓ ***Matrice creuse :*** Une solution consiste à exploiter la technique basée sur le contenu [Lang,1995] [Krulwich et Burkey, 1996] [Billsus et Pazzani, 2000]. Cette technique est utilisée tant que les notes sur un item ne sont pas suffisamment disponibles. Quand un nouvel

item est introduit, la technique basée sur le contenu évalue la similarité de son contenu avec les items disponibles afin de l'impliquer au processus de recommandation. Néanmoins, l'utilisation de la technique basée sur le contenu engendre un manque de diversité des recommandations, ce qui entrave la performance du système de recommandation. [Mobasher, 2004] proposent une mesure de similarité qui combine la similarité sémantique entre documents avec celle basée sur les votes .Les travaux de [Wang, 2006] montrent que la fusion des deux similarités d'utilisateurs et des ressources permettent d'améliorer la précision de la prédiction, en limitant l'effet des matrices creuses. [Park, 2006] ont proposé l'utilisation des techniques à base d'agents « agents d'évaluations automatiques » pour résoudre le problème de démarrage à froid lorsque la matrice utilisateurs/documents est creuse,

Certaines approches utilisent la similarité sémantique pour compléter la matrice creuse des évaluations statistiques comme la technique des évaluations implicites [Jin, 2003] [Ziegler, 2004] [Melville, 2002]

✓ ***le passage à l'échelle*** *:* Plusieurs approches peuvent être adoptées pour résoudre cette problématique, parmi les quelles :

• l'utilisation de modèles (comme le Clustering), Pour améliorer la qualité du clusters construit. [Cantador, 2006] proposent une approche de Clustering sémantique. Ils sont utilisés des ontologies de base qui représente les profils des utilisateurs pour construire les clusters par l'exploitation des préférences communes entre les utilisateurs.

• Calculs faits en arrière plan.

• La réduction de dimension, par la décomposition des matrices ou par clustering.

• Les modèles distribués, le traitement est partagé sur plusieurs ressources physiques.

4. Conclusion :

Malgré que l'utilisation de web sémantique et surtout les ontologies a nécessité la construction préalable d'une ontologie relative au domaine de connaissance et d'autre part, la construction d'une ontologie est un processus complexe et coûteux, il est clair, en présentant l'intégration du web sémantique dans le filtrage collaboratif,que son utilisation est avantageu par rapport à d'autres approches, car il offre la possibilité d'effectuer des raisonnements et la définition de règles à l'aide des ontologies bien définies. En plus, il donne des solutions aux problèmes de :

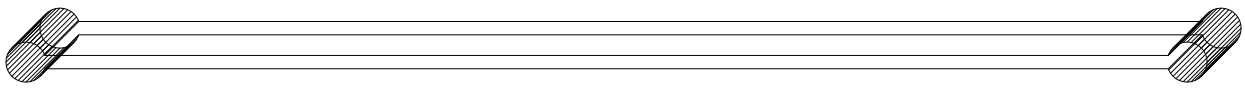
✓démarrage à froid

✓Matrice creuse

✓passage à l'échelle

Malgré toutes les idées proposées et les système réalisés ,le problème de la matrice creuse n'est pas résolus à 100% et en plus un autre problème de l'effet entonnoir s'ajoute où la majorité des systèmes actuels ne permettent pas dans certains cas de prendre en compte les documents d'un nouvel axe de recherche pour les différentes communautés -, et toutes les fonctions de prédiction se basent sur les données disponibles dans la matrice utilisateur /item, quand il y a un manque de données, la modélisation des utilisateurs et des items devient de plus en plus difficile et complexe.

Chapitre 4



Approche proposée et implémentation

1. Introduction :

Dans ce chapitre, nous proposons une approche de filtrage d'information hybride, qui se base sur l'algorithme basé mémoire afin de l'améliorer et de réduire les effets des problèmes présentés en-dessus .Nous proposons l'intégration du principe de l'algorithme basé modèle «le regroupement des utilisateurs similaires » et la description sémantique des documents à recommander « ontologie de domaine ». De ce fait, nous proposons un schéma où :

✓ *La formation des communautés :* Dans le but de réduire la taille de la matrice utilisateurs/items, nous proposons l'utilisation de l'ontologie de domaine afin de définie la liste de toutes les communautés disponibles dans le système, qui sont envoyées aux utilisateurs pour choisir la communauté préférée, le seul responsable sur le choix est l'utilisateur. Afin d'enrichir ces communautés, un délégué « représentant » est défini pour chacune et dans chaque communauté le nombre d'utilisateurs est augmenté où nous ajoutons les représentants.

✓ *Production des recommandations :* nous nous basons dans les calcul sur les fonctions de l'algorithme basé mémoire.

✓*Récupération d'evaluation :* C'est l'étape qui consiste à regrouper les jugements des utilisateurs portés sur les documents recommandés. Ces jugements aident le système pour former les communautés. Afin de les capturer nous avons défini un algorithme sémantique qui s'exécute en arrière plan, de deux manières : explicite (l'utilisateur donne une évaluation manuelle) et implicite (l'évaluation est récupérée automatiquement). Dans ce dernier l'evaluation de l'utilisateur actif « UtilA » sur l'item actif « ItemA » est calculée à partir des évaluations des items similaires « similarité sémantique » au « ItemA » donnée par les utilisateurs similaires de « UtilA » et la valeur trouvée normalisée par l'évaluation moyenne de l'ItemA. La similarité sémantique est calculée à l'aide d'une fonction proposée basée sur le nombre de recommandations de chaque item.

Afin de détailler notre approche, nous donnons une description détaillée de chacun de ces points.

2. Architecture générale :

L'idée de notre proposition (figure 4.1) est l'adoption d'un système hybride qui combine les avantages des systèmes basé modèle « création des communautés » et les systèmes basé-mémoire « la diversité des recommandations » et du web sémantique « les avantages des ontologies » afin de réduire les effets des problèmes présentés précédemment.

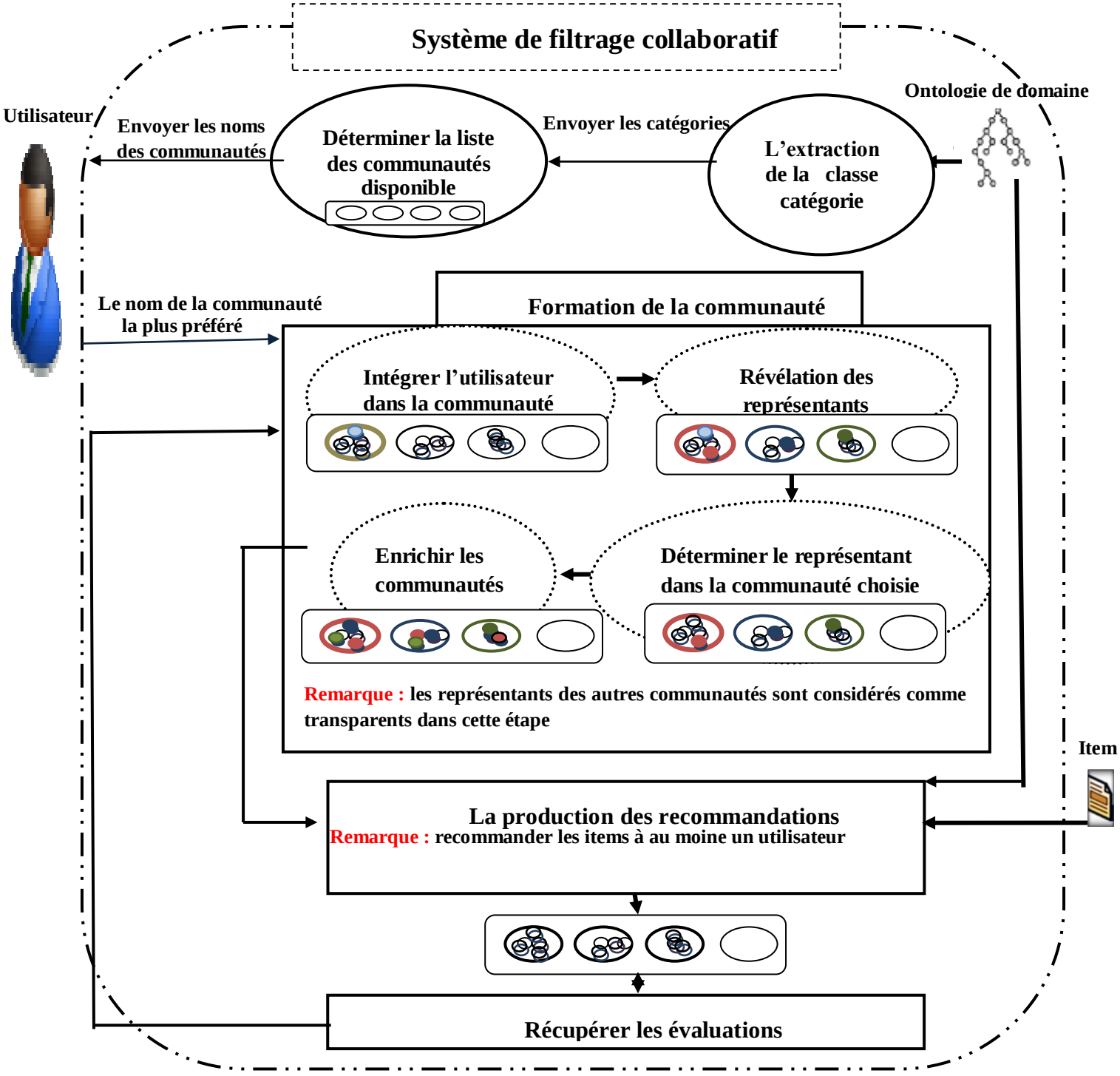


Figure 4.1 : L'architecture générale de notre approche

2.1. Détermination de la liste des communautés disponibles :

Dans cette étape le système forme depuis l'ontologie d'items en particulier la classe catégories d'items, la liste des noms des communautés disponibles dans le système. La figure

4.2 présente le processus de formation où le système manipule l'ontologie et extrait les classes de catégorie les plus élémentaire c.à.d. les classes qui ne contiennent pas des sous classes.

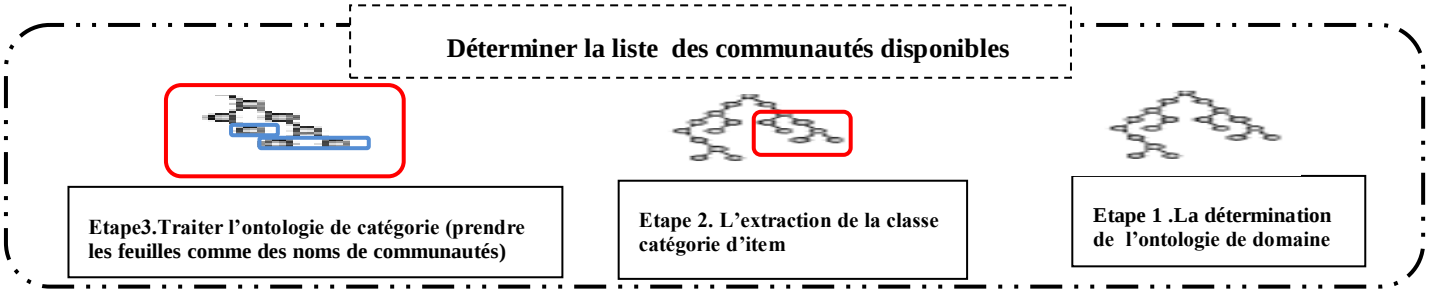


Figure 4.2 : La détermination de la liste des communautés disponibles

Liste $_{communautés}$ = **{**($C_{catégorie1}$, (P_{I1} ….P_{IN})),… ($C_{catégorie\ I}$, (P_{I1} ….P_{IN})),…., ($C_{catégorieN}$,(P_{I1} ….P_{IN}))**}**

$C_{catégorie\ I}$ =Nom de la communauté I

P_{IJ} : Propriété J de la communauté I

Par la suite le système à l'arrivée d'un nouvel utilisateur, il envoie la liste des noms des communautés disponible, pour que l'utilisateur choisisse la communauté la plus préférée.

2.2. Formation de la communauté :

Dans cette phase on a défini une nouvelle notion « les représentants » qui sont les délégués des utilisateurs, pour chaque communauté, comme c'est illustré sur la figure 4.3 :

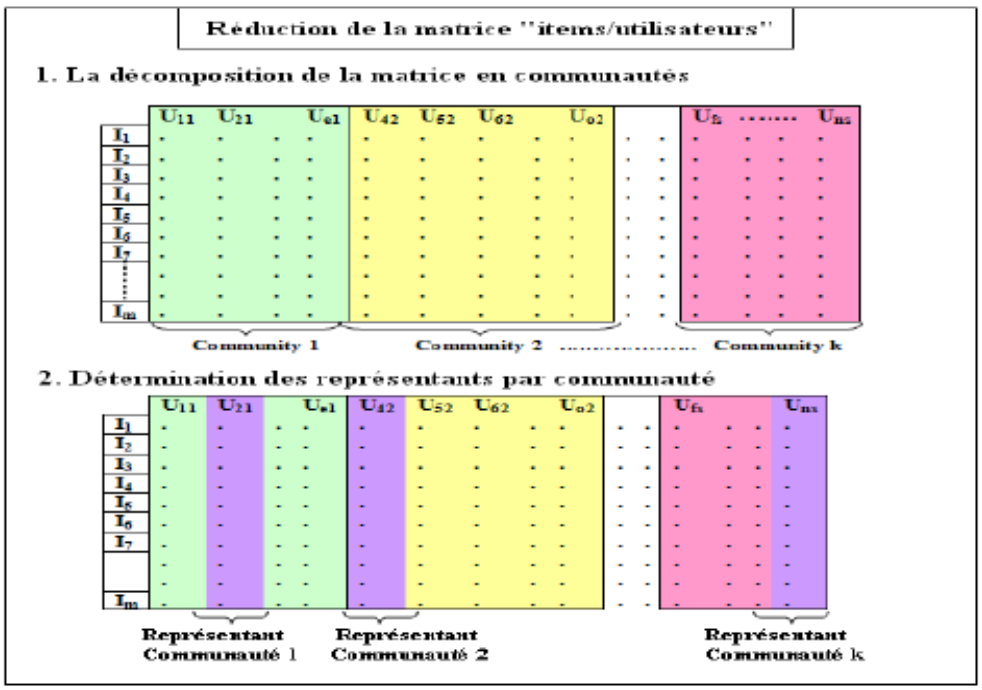
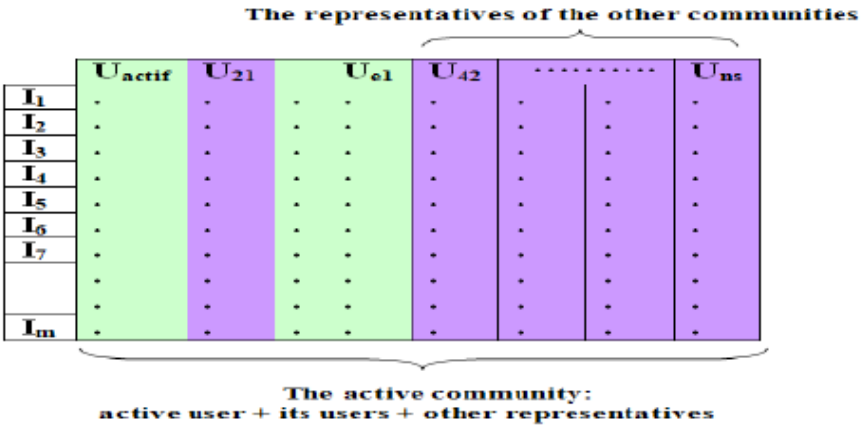


Figure 4.3 : Le processus de formation des communautés

Ensuite au lieu d'utiliser la matrice utilisateurs/items entière, chaque utilisateur à *une communauté* bien définie et les calculs sont effectués sur seulement leur communauté qui doivent être enrichies par les représentants afin de régler le problème de l'effet entonnoir :



Notre processus de formation se déroule comme suit :

✓ *Intégrer l'utilisateur dans la communauté qu'il a choisi*.

✓ *Révélation des représentants :* Seulement dans les communautés qui sont des utilisateurs préalablement, le système extrait pour chaque communauté leur pur représentant.

✓ *Déterminer le représentant dans la communauté choisie* **:** Si dans la communauté on a un seul utilisateur « i » alors « i » est le représentant, sinon si on a seulement des nouveaux utilisateurs (c.à.d. Les utilisateurs n'ont pas des évaluations) alors le choix du représentant est aléatoire, sinon on calcule le vote moyen de tous les utilisateurs de la communauté sauf pour les nouveaux utilisateurs par la fonction de [Melville ,2002] qui représente le centre de gravité :

$$\text{Le centre de gravité} = \frac{\sum_{user=1}^{m} \left(\frac{\sum_{item=1}^{n} vote(user, item)}{n} \right)}{m}$$

Et on calcule pour chaque utilisateur le vote moyen :

$$\text{Vote moyen} = \frac{\sum_{item=1}^{n} vote(user, item)}{n}$$

Le délégué représente l'utilisateur qui à le vote moyen le plus proche du centre de gravité

✓ *Enrichir les communautés* **:** Ces représentants sont intégrés dans chaque communauté du système. Donc dans chaque communauté on à :

$C_{\text{catégorieI}} = \{(U_1, U_2, \ldots, U_J, \text{Rep}_{\text{catégorieI}}, \text{Rep}_{H=\{1..N\}-I)}\}$

« Les utilisateurs (U_J), le pur représentant $\text{Rep}_{\text{catégorie I}}$ *et les représentants des « N » autres*

communautés Rep $_{H=\{1..N\}-I)}$ »

Remarque : Malgré qu'on a dans un système qui apparait comme basé modèle mais notre idée est d'améliorer l'algorithme basé-mémoire par la réduction de la matrice utilisateurs / items.

2.3. Production des recommandations :

Ce processus est déclenché pour deux raisons :

✓*Réception d'un nouvel item* **:** Est un problème pour le filtrage collaboratif, ce document n'est pas encore évalué, et les objets à recommander ne sont décrits que par les évaluations fournies par les utilisateurs, pour cela on va profiter des informations sémantiques de cet item afin de le recommander, où le système compare les propriétés sémantiques des items avec ceux des communautés (figure 4.4) par une fonction booléenne et envoyer l'item aux communautés (un ou plus) qui ont des propriétés en commun. Dans le cas où le nombre de communautés sélectionné égale à 0 le système envoi l'item à toutes les communautés.

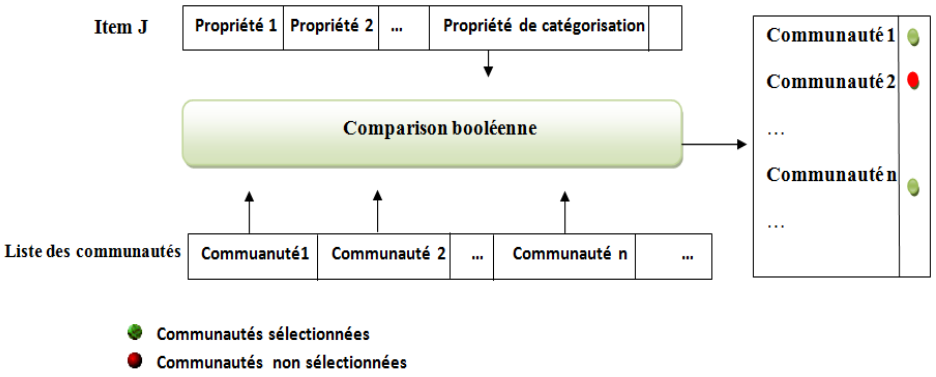


Figure 4.4 : La production des recommandations

Une fois l'item est évalué par les utilisateurs des communautés sélectionnées, si l'évaluation et positive ou nulle (>à un seuil), l'item est envoyé aux communautés non sélectionnées.

✓*Reformation des communautés pour un nouvel utilisateur* **:** Pour calculer la prédiction de l'utilisateur actif «a » sur l'item « j », on utilise la fonction de prédiction de l'algorithme basé-mémoire, en appliquant sur la communauté enrichie qui contient l'utilisateur «a » et les représentants « centres de gravité » des autres communautés :

$$p_{a,j} = \overline{v_a} + k \sum_{i=1}^{n} w(a,i)(v_{i,j} - \overline{v_i}) \quad Avec \quad \overline{v_i} = \frac{1}{|I_i|} \sum_{j \in I_i} v_{i,j}$$

Où $\overline{v_i}$: L'évaluation moyenne de l'utilisateur « i »

I_i : L'ensemble des items évalués par l'utilisateur « i »

K : Coefficient de normalisation

n : Les utilisateurs similaires de « a » plus et les représentants.

W (a, i) : La similarité calculée entre l'utilisateur actif « a » et les autres utilisateurs « i » de la mêmes communauté :

Si (« i » n'est pas un représentant) Alors W (a, i) est calculée par la fonction de corrélation

$$w(a,i) = \frac{\sum_j (v_{a,j} - \overline{v_a})(v_{i,j} - \overline{v_i})}{\sqrt{\sum_j (v_{a,j} - \overline{v_a})^2 \sum_j (v_{i,j} - \overline{v_i})^2}}$$

Sinon W (a, i) = M {un constant}

2.4. Récupération des évaluations:

C'est une étape très sensible, elle influence les deux autres processus. L'algorithme 1 présente les étapes suivies afin de récupérer les évaluations des utilisateurs en arrière plan, de deux manières explicite (l'utilisateur donne une évaluation) et implicite(l'évaluation est récupérée par l'algorithme) .Dans ce cas l'evaluation de l'utilisateur UtilA sur l'item ItemA est calculée à partir des evaluation des items similaires « similarité sémantique » à « ItemA » donnée par les utilisateurs similaire de « UtilA » et la valeur trouvée normalisée par l'évaluation moyenne de l'ItemA. La similarité sémantique est calculée à l'aide d'une fonction proposée basée sur le nombre de recommandation de chaque item.

Algorithme 1

Variables
UserA: L'utilisateur atif,
ItemA: L'item recommandé,
v[i,j]: La matrices des votes (évaluations),
θ: Seuil (0≤θ≤1),
Nb_recmd_item[j]: compte le nombre de fois où l'item "j" a été recommandé

Begin
//l'utilisateur a donné une évaluation à l'item et on remplit la matrice de //votes
if(Evaluation (UserA, ItemA)>0) then
v[UserA,ItemA] = Evaluation (UserA, ItemA)
else
Begin
//déterminer le nombre de fois de recommandation de l'item qui n'ont //pas d'évaluations
for (item = 1 to m) do
Recommend[item]= Nb_recmd_item[item];

//calculer la similarité entre l'item actif et les autres items
for (item = 1 to m) do
if (Recommend[ItemA] > Recommend[item]) then

$$Similarity[item, ItemA] = \frac{\text{Re}\,commend[item]}{\text{Re}\,commend[ItemA]}$$

else

$$Similarity[item, ItemA] = \frac{\text{Re}\,commend[ItemA]}{\text{Re}\,commend[item]}$$

//déterminer les items similaires de ItemA
Nb_item_sim=0;
for (item =1 to m) do
if (Similarity[item, ItemA] ≥ θ) then Nb_item_sim++;

//calculer le vote moyen de ItemA (Normalisation)

$$AV(ItemA) = \frac{\sum_{user=1}^{n} v(user, ItemA)}{n}$$

// calculer la similarité entre users "i" et UserA "a"
for (user = 1 to n) do

$$w(a,i) = \frac{\sum_j (v_{a,j} - \overline{v_a})(v_{i,j} - \overline{v_i})}{\sqrt{\sum_j (v_{a,j} - \overline{v_a})^2 \sum_j (v_{i,j} - \overline{v_i})^2}}$$

//déterminer les utilisateurs similaires de UserA
Nb_user_sim=0;
for (user = 1 to n) do
if (w[item, ItemA] ≥ θ) then Nb_user_sim++;

//calculer le vote moyen des items similaires à ItemA
//pour tous les utilisateurs similaires de UserA

$$vote\_items\_sim[i] = \frac{\sum_{j=1}^{Nb\_user\_sim} v(j,i)}{Nb\_user\_sim}$$

$$Vote\_Moyen(ItemA) = \frac{\sum_{j=1}^{Nb\_item\_sim} vote\_items\_sim(j)}{Nb\_item\_sim}$$

//calculer le vote de UserA pour ItemA

$$v(UserA, ItemA) = \frac{AV(ItemA) + Vote\_Moyen(ItemA)}{2}$$

End
End

Ainsi pour un utilisateur « i », si la majorité des évaluations récupérées pour différents items sont négative (évaluations négatives>évaluations positives) alors nous concluons que la communauté de cet utilisateur est mal choisie ; dans ce cas-ci le système calcule la similarité

entre l'utilisateur « i » et les représentants des autres communautés et choisit la communauté qui inclut le représentant le plus semblable (la valeur la plus élevée).

3. Implémentation et résultats :

Dans cette section nous présentons notre application afin de tester notre approche. Pour ce faire, nous avons choisi l'utilisation du jeu de données MovieLens.

Nous avons eu recours à l'implémentation en Java de tous les modules .Ensuite nous avons procédé une évaluation, afin d'étudier le comportement de système en variant le nombre d'évaluations donné par les utilisateurs.

Une discussion des résultats de chaque module est enfin présentée.

3.1. Le jeu de données :

3.1.1. Données évaluations :

Afin de développer un programme pour appliquer et tester notre approche proposée du filtrage collaboratif, nous avons utilisé la base MovieLens (http://www.movielens.org/). Cette base contient les informations de 1.682 films, 943 utilisateurs et 100.000 votes. Chaque utilisateur a voté au moins 20 films. La valeur d'un vote est entre 1 et 5. Les fichiers principaux de la base MovieLens sont les suivants : u.data, u.item, u.genre, u.user.

Le fichier u.data contient 100.000 votes créés par 943 utilisateurs sur 1.682 films (items). C'est une liste dont les champs sont séparés par des tabs comme suit :

user id | item id | rating

Les informations des films (items) sont contenues dans le fichier u.item de façon comme suit :

movie id | movie title | release date | video release date | IMDb URL | unknown | Action | Adventure | Animation | Children's | Comedy | Crime | Documentary | Drama | Fantasy | Film-Noir | Horror | Musical | Mystery | Romance | Sci-Fi | Thriller | War | Western

Les derniers 19 champs sont des genres, un 1 indique que le film est à ce genre, un 0 indique que non; les films peuvent être avoir plusieurs genres en même temps. Les « movie id » sont ceux utilisés dans la liste d'u.data.

```
1|Toy Story (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Toy%20Story%20(1995)|0|0|0|1|1|1|0|0|0|0|0|0|0|0|0|0|0|0
2|GoldenEye (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?GoldenEye%20(1995)|0|1|1|0|0|0|0|0|0|0|0|0|0|0|0|1|0|0
3|Four Rooms (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Four%20Rooms%20(1995)|0|0|0|0|0|0|0|0|0|0|0|0|0|0|0|1|0|0
4|Get Shorty (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Get%20Shorty%20(1995)|0|1|0|0|0|1|0|0|1|0|0|0|0|0|0|0|0|0
5|Copycat (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Copycat%20(1995)|0|0|0|0|0|0|1|0|1|0|0|0|0|0|0|0|1|0|0
6|Shanghai Triad (Yao a yao yao dao waipo qiao) (1995)|01-Jan-1995||http://us.imdb.com/Title?Yao+a+yao+yao+dao+waipo+qiao+(1995)|0|0|0|0|0|0|0|0|0|1|0|0|0|0|0|0|0|0|0|0|0
7|Twelve Monkeys (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Twelve%20Monkeys%20(1995)|0|0|0|0|0|0|0|1|0|0|0|0|0|0|1|0|0|0
8|Babe (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Babe%20(1995)|0|0|0|0|1|1|0|0|1|0|0|0|0|0|0|0|0|0|0
9|Dead Man Walking (1995)|01-Jan-1995||http://us.imdb.com/M/title-exact?Dead%20Man%20Walking%20(1995)|0|0|0|0|0|0|0|0|1|0|0|0|0|0|0|0|0|0|0
10|Richard III (1995)|22-Jan-1996||http://us.imdb.com/M/title-exact?Richard%20III%20(1995)|0|0|0|0|0|0|0|0|1|0|0|0|0|0|0|0|1|0
.....
```

La liste des genres est contenue dans le fichier u.genre.

```
unknown|0
Action|1
Adventure|2
Animation|3
Children's|4
Comedy|5
.....
```

Le fichier u.user contient des informations démographiques des utilisateurs. Les champs sont séparés par des tabs.

user id	age	gender	occupation	zip code

```
1|24|M|technician|85711
2|53|F|other|94043
3|23|M|writer|32067
4|24|M|technician|43537
5|33|F|other|15213
6|42|M|executive|98101
7|57|M|administrator|91344
8|36|M|administrator|05201
9|29|M|student|01002
10|53|M|lawyer|90703
.....
```

3.1.2. Données sémantiques :

Afin d'évaluer la solution proposée, il est nécessaire de pouvoir représenter la dimension sémantique des items. Pour ce faire, nous avons utilisé la classification hiérarchique des films de MovieLens de :

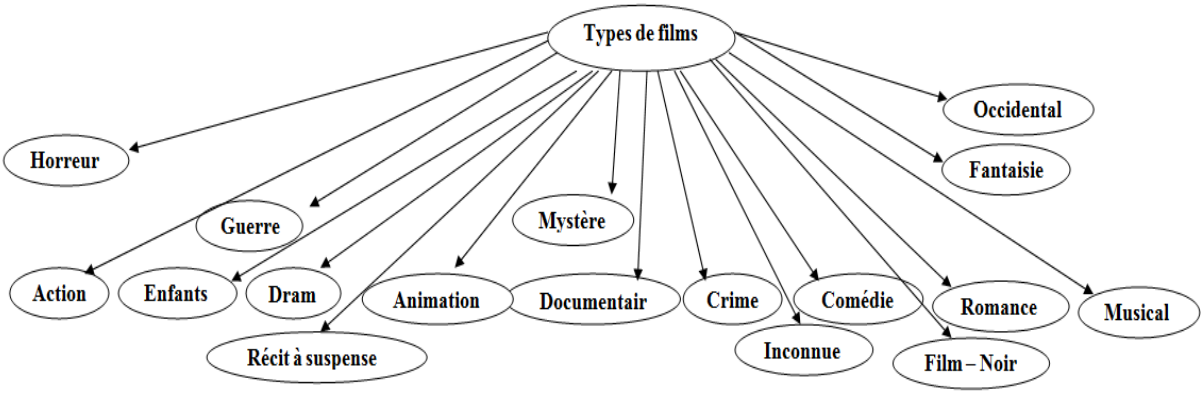


Figure 4.5 : La classification hiérarchique des films de MovieLens

Où le fichier u.item visualise les genres de chaque film :

titre	daterealisat	datevedio	URL	Action	Adventure	Animation	Comidie	Crime	Documentai	Drame	Enfants	Fantaisie	Film - Noir
Three Colors: \	00/00/1994	01/01/1994	http://us.imdt	0	0	0	0	0	0	1	0	0	0
Grand Day Out	00/00/1992	01/01/1992	http://us.imdt	0	0	1	1	0	0	0	0	0	0
Desperado	00/00/1995	01/01/1995	|http://us.imc	1	0	0	0	0	0	0	0	0	0
Glengarry Gler	00/00/1992	01/01/1992	http://us.imdt	0	0	0	0	0	0	1	0	0	0
Angels and Ins	00/00/1995	01/01/1995	http://us.imdt	0	0	0	0	0	0	1	0	0	0
Groundhog Da	00/00/1993	01/01/1993	http://us.imdt	0	0	0	1	0	0	0	0	0	0
Delicatessen	00/00/1991	01/01/1991	http://us.imdt	0	0	0	1	0	0	0	0	0	0
Hunt for Red C	00/00/1990	01/01/1990	http://us.imdt	0	1	0	0	0	0	0	0	0	0
Dirty Dancing	00/00/1987	01/01/1987	http://us.imdt	0	0	0	0	0	0	0	0	0	0
Good, The Bad	00/00/1996	01/01/1996	http://us.imdt	1	0	0	0	0	0	0	0	0	0

3.2. Prototype d'implémentation :

A l'aide du fichier u.item de Movielens et la classification hiérarchique présenté au dessus, La liste des communautés disponibles dans notre système est la suivante : Action, Adventure, Animation, Comédie, Crime, Documentaire, Drame, Enfants, Fantaisie, Film – Noir, Guerre, Horreur, Inconnue, Musical, Mystère, Occidental, Récit à suspense, Romance

Par l'analyse de fichier u.data qui contient les votes créés par les utilisateurs sur les films,

La plupart des utilisateurs évaluent avec des notes, dans l'ordre décroissant, de 5,4 et 3

- 23 % par des notes de 5
- 35 % des évaluations effectuées sont représentées par des notes de 4
- 26 % par des notes de 3

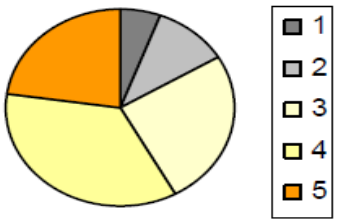


Figure 4.6 : La répartition des évaluations

Ce qui signifie que les utilisateurs n'évaluent que rarement les ressources qu'ils considèrent comme non intéressantes et se focalisent plutôt sur les ressources qui préfèrent ce qui facilite leur distribution sur les communautés où chaque utilisateur est intégré dans la communauté de genre préféré d'item : *Le nombre d'utilisateur dans chaque communauté est entre le 1 et 100*

Chaque communauté contient un certain nombre d'utilisateurs. Ainsi, nous avons obtenu 18 matrices. Chaque matrice a 1682 lignes (items) et le nombre de colonnes est réduit selon le nombre d'utilisateurs de la communauté à qui on ajoute 17 colonnes (représentants). Ainsi, nous avons les matrices suivantes : Action (100 utilisateurs), aventure (50), (animation, 40), (comédie, 100), (crime, 60), (documentaire, 10), (drame, 100), (enfants, 43), (imagination, 40), (film - Noir, 10), (guerre, 70), (horreur, 100), (inconnu, 1), (musical, 50), (mystère,19), (Occidental, 25), (suspens, 25), et (Romance,100). Chaque matrice est utilisée pour calculer l'évaluation de prévision par notre approche.

3.4. Résultats et discussion :

Dans nos essais, nous choisissons aléatoirement un item (film), de l'ensemble de données ; nous choisissons également aléatoirement 10 utilisateurs qui ont évalué (voté) ce film. Pour chaque utilisateur, nous calculons la valeur de prévision par notre méthode et par l'algorithme basé-mémoire. Nous comparons le vrai vote qui est dans la base de données par rapport aux deux méthodes. Nous augmentons le nombre d'articles, et nous faisons la même opération. Pour chaque groupe d'articles, nous choisissons 10 utilisateurs qui ont évalué ces films et calculons la prévision moyenne avec ces 2 méthodes.

La comparaison entre la prévision et le vrai vote est faite par l'erreur absolue moyenne (MAE) qui correspond à l'erreur absolue moyenne entre le vote de l'utilisateur « ri » et la prévision du vote « pi ».

$$MAE = \frac{\sum_{i=1}^{10} |p_i - r_i|}{10}$$

3.4.1. Le calcul de prédiction :

Pour calculer la prédiction de l'utilisateur actif «a » sur l'item « j », on utilise la fonction de prédiction de l'algorithme basé mémoire, en appliquant sur la communauté enrichie qui contient l'utilisateur «a » et les représentants « centres de gravité » des autres communautés:

$$p_{a,j} = \overline{v_a} + k \sum_{i=1}^{n} w(a,i)(v_{i,j} - \overline{v_i}) \quad Avec \quad \overline{v_i} = \frac{1}{|I_i|} \sum_{j \in I_i} v_{i,j}$$

Où W (a, i) est calculée par la fonction de corrélation *si* « i » n'est pas un représentant

$$w(a,i) = \frac{\sum_j (v_{a,j} - \overline{v_a})(v_{i,j} - \overline{v_i})}{\sqrt{\sum_j (v_{a,j} - \overline{v_a})^2 \sum_j (v_{i,j} - \overline{v_i})^2}}$$

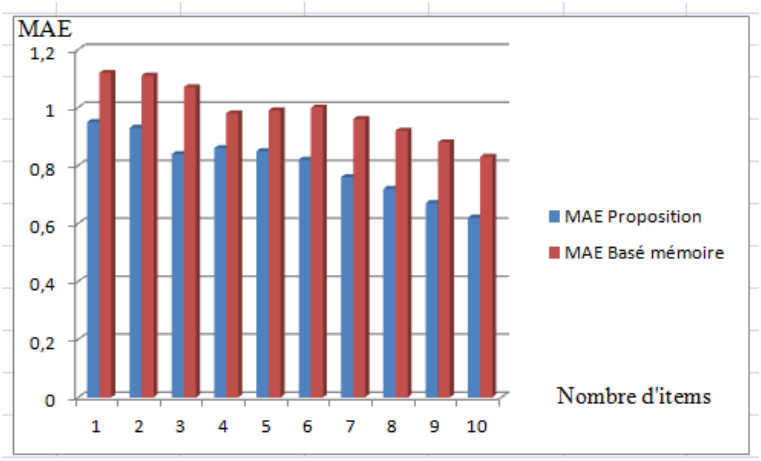
Sinon W (a, i) = *0.6*

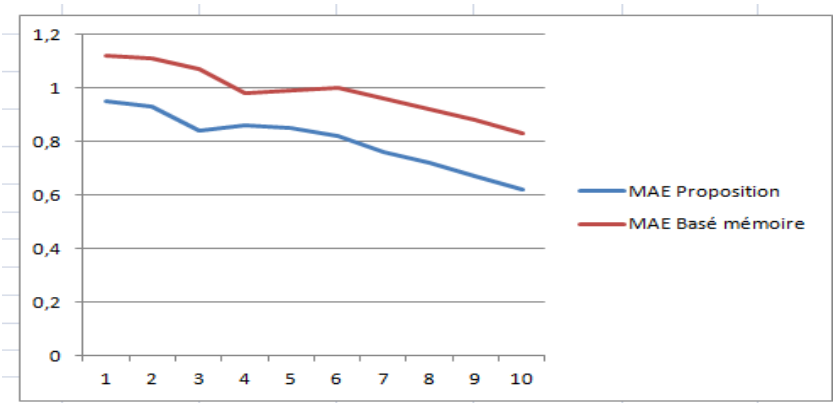
Le résultat trouvé est le suivant :

Nombre d'items	MAE Proposition	MAE Basé mémoire
1	0,95	1,12
2	0,93	1,11
3	0,84	1,07
4	0,86	0,98
5	0,85	0,99
6	0,82	1
7	0,76	0,96
8	0,72	0,92
9	0,67	0,88
10	0,62	0,83

MAE Proposition	MAE Basé mémoire	Couverture
8,02	9,86	100%

Les schémas correspondent au MAE et de la prédiction de l'algorithme basé mémoire et notre approche proposée :





Discussion :

Après l'analyse des résultats trouvées La précision de l'approche proposée est plus grande que celle de l'algorithme basé mémoire, où le comportement est différent, donc comme notre proposition basée sur l'utilisation des données sémantique afin de réduire la taille de la matrice utilisateur /item et l'intégration des délégués dans les différentes communautés, nous avons gagné un temps de calcul sans perte de données.

Nous pouvons donc conclure qu'avec l'utilisation de cette approche avec les données sémantiques des items comme moyen de création des communautés qui englobe tous les utilisateurs qui ont des tendances similaires de façon manuelles ,et les enrichir par les représentants , peut arriver à un degré important de précision où la différence entre la valeur de prédiction de notre approche et l'algorithme basé-mémoire est remarquable.

Nous pouvons dire qu'on a réduit l'effet entonnoir par l'intégration des représentants et au passage à l'échelle par la réduction de la taille de la matrice utilisateurs /items.

3.4.2. La récupération des évaluations :

C'est une étape très sensible. Elle est effectuée en arrière plan pour gagner le temps et minimiser le passage à l'échelle , de deux manières :explicite (l'utilisateur donne une

évaluation) et implicite(l'évaluation est récupérée par l'algorithme proposé) .Dans le deuxième cas l'evaluation de l'utilisateur « UtilA » sur l'item « ItemA » est calcule a partir des évaluations des items similaires « similarité sémantique proposée » à « ItemA » donnée par les utilisateurs similaires de « UtilA » et la valeur trouvée normalisée par l'évaluation moyenne de l'ItemA. La similarité sémantique est calculée à l'aide d'un algorithme proposé basé sur le nombre de recommandation de chaque item.

✓**Similarité sémantique :**

Les similarités calculé par l'algorithme proposé, avec un seuil égale à 0.5 est comparait par celle de la fonction de Cosinus :

$$sim_{cosine}(\vec{x}, \vec{y}) = \frac{\vec{x} \bullet \vec{y}}{|\vec{x}| \times |\vec{y}|}$$

Le résultat trouvé est le suivant :

Item 1	Item 2	Similarité sémantique proposée	Tesrte de similarité	La fonction de cosinus	Tesrte de similarité
1	1	1	Oui	1	Oui
1	11	0,5	Oui	0,42	Non
1	12	0,57	Oui	0,41	Non
2	119	0,09	Non	0,3	Non
2	131	0,92	Oui	0,02	Non
2	125	0,42	Non	0,27	Non
2	1129	0	Non	0	Non
2	1127	0,09	Non	0	Non
2	285	1	Oui	0	Non

Discussion :

Après l'analyse des résultats trouvés on a remarqué que :

✓Les valeurs de similarité trouvées sont différentes

✓ Presque tous les items qui sont similaires après l'application de la fonction de Cosinus sont aussi similaires par notre fonction

✓Il ya des items qui ne sont pas similaires après l'application de la fonction de Cosinus qu'ils sont similaires par notre fonction

Donc nous pouvons conclure qu'avec la définition de cet algorithme nous avons amélioré la similarité entre les items où le nombre des items similaires a augmenté, ce que permet de minimiser le l'effet du problème de manque de données dans le cas de la matrice creuse.

✓L'algorithme de récupération des évaluations :

La valeur de seuil de similarité utilisé est 0.5 où :

Si (la similarité entre deux items >= 0.5) Alors les deux items sont similaires

Sinon les deux items ne sont pas similaires

Le résultat trouvé est le suivant :

user	Item	Le vote réel	Le vote prédit	MAE
a	1	4	3,5	0,5
b	1	4	3,5	0,5
c	1	4	4,5	0,5
d	1	5	4,5	0,5
e	2	3	3,5	0,5
f	2	3	3,5	0,5
g	2	3	3,5	0,5
h	2	3	3,5	0,5

Discussion :

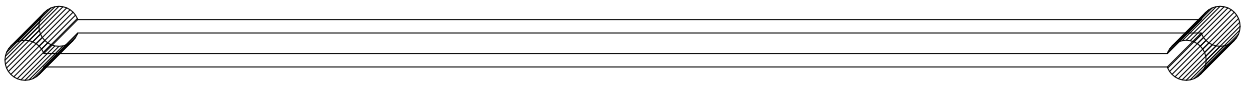
Après l'analyse des résultats trouvés nous avons remarqué que :

✓ Presque pour chaque item nous avons un nombre fixe des évaluations qui sont répétés

✓Presque tous les utilisateurs de la même communauté on la même valeur d'évaluation pour un item donné

✓Pour les évaluations trouvées la valeur de MAE et entre (0.5 et 1.5)

Donc nous pouvons conclure qu'avec cet algorithme on peut calculer des évaluations à partir des évaluations des utilisateurs similaires de l'utilisateur actif sur les items similaires de l'item actif, mais dans le cas où les évaluations utilisées dans les calculs sont de deux niveaux différents (le niveau 1 : évaluation 5, 4et 3 et le niveau 2 : évaluation 1 et2) le niveau qui à le nombre petit d'évaluations est négligeable.



Conclusion générale

Conclusion et perspectives :

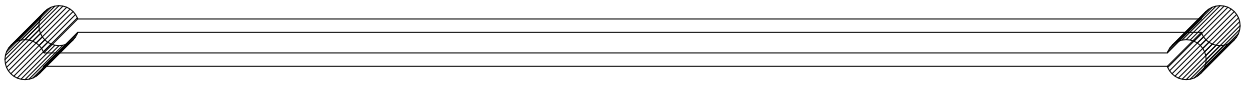
Dans ce travail, nous avons présenté une vue d'ensemble sur le filtrage collaboratif et ses inconvénients. Nous avons essayé de combiner le Web sémantique avec l'algorithme basé-mémoire pour réduire la matrice « items/utilisateurs ». On a proposé une nouvelle notion qui est le « représentant ». Ce dernier représente un ensemble d'utilisateurs et réduit donc la matrice « items/utilisateurs ». Nous pouvons dire que les données sémantiques et les représentants ont apporté une amélioration significative à la qualité des systèmes de recommandation. Le cadre structural et un algorithme ont été développés et mis en application. Quelques résultats appropriés ont été présentés à la fin de ce travail. Pour examiner notre approche, nous l'avons appliquée sur l'ensemble de données de MovieLens. Une comparaison entre la méthode implantée en mémoire et notre approche a été donnée.

Par conséquent, cette approche semble résoudre en partie l'effet d'entonnoir et le passage à l'échelle et en réduisant la taille de la matrice utilisateurs/items par l'intégration des représentants et les calculs en arrière plan.

Nous sommes sortis avec des conclusions positives sur le fait que les données sémantiques apportent une amélioration notable à la qualité des systèmes de recommandation basée mémoire, en améliorant la précision, où les données ne sont pas perdues.

Nous pouvons dire que cette approche semble résoudre en partie la matrice creuse, excepté dans le cas où l'item actif ne contient aucun item similaire ou si tous les utilisateurs similaire à l'utilisateur actif ont des évaluations nulles.

Comme cette idée est nouvelle, les perspectives de recherche sont nombreuses, le premier axe est la recherche des différents standards sémantiques du web 2.0 qui pourraient améliorer la description des communautés, une autre perspective est la détermination du nombre des représentants efficaces pour chaque communauté afin d'améliorer la précision et intégrer un autre aspect dans l'algorithme de récupération des évaluations afin d'avoir des valeurs exacte.



Références

Référence :

[Abbes,1999] : R.Abbes. « Le filtrage des informations » .1999.

[Aha, 1991] :AHA, D. W.; KIBLER, D.; ALBERT, M.K., Instance-Based Learning Algorithms, Machine Learning, 6, pp. 37-66, 1991.

[Amokrane,2007]:B. Amokrane. « L'usage des concepts du web sémantique dans le filtrage d'information collaboratif »,2007.

[An, 2006] :T.An . COCoFil2 : « Un nouveau système de filtrage collaboratif basé sur le modèle des espaces de communautés » ,2006 .

[Anand et Mobasher, 2005] : Anand, S. et Mobasher, B. (2005). « Intelligent techniques for web personalization". Lecture Notes in Artificial Intelligence, 3169:1–36.

[Belkin et Croft, 1992] : N.Belkin,et W.Croft. « Information retrieval and information filtering : two sides of the same coin»,Communications of the ACM, 35(12), 1992.

[Bisiaux,2003] :C. Bisiaux , « Expérience sur l'utilisation de conjonctions de termes et la prise en compte des dépendances entre termes d'indexation dans le cadre du filtrage adaptatif » ,2003.

[Boughanem,2001] :M.BOUGHANEM , M . TMAR . « Filtrage d'information par combinaison d'un profil positif et profil négatif » , IRIT/SIG , Compus Univ Toulouse III , Université de Nantes:Paris X , 3° congrès du chapitre français di l'ISKO, p209:217 ,2001 .

[Breese, 1998]:J. Breese , D.Heckerman , C.Kadie . « Empirical Analysis of Predictive Algorithms for Collaborative Filtering », Proceedings of the 14th Conference on Uncertainty In Artificial Intelligence (UAI'98), Wisconsin, USA, p. 43-52,1998.

[Berrut, 2003] :C.Berrut ,N.Denos . «Filtrage collaboratif», Assistance intelligente à la recherche d'informations, Hermes - Lavoisier, chapter 8, pp30, 2003.

[Billsus et Pazzani, 2000] : Seventh International Conference BILLSUS, D.; PAZZANI, M., A Hybrid User Model for News Story Classification, Proceedings of the on User Modeling (UM '99), Banff, Canada.

[Burke ,2002] :R.Burke . « Hybrid Recommender Systems: Survey and Experiments », Journal of Personalization Research, User Modeling and User-Adapted Interaction, vol. 12 (4), , Kluwer Academic Publishers, p. 331-370, 2002.

[Castagnos ,2006] : Sylvain Castagnos and Anne Boyer. A Client/Server User-Based Collaborative Filtering Algorithm: Model and Implementation. 17th European Conference on Artificial Intelligence (ECAI 2006), in the 4th Prestigious Applications of Intelligent Systems special section (PAIS).Riva del Garda, Italy, August 2006.

[Chee, 2001]: Sonny Han Seng Chee, Jiawei Han, Ke Wang. RecTree: An Efficient Collaborative Filtering Method. In Proceeding 2001 Int. Conf. On Data Warehouse and Knowledge Discovery (DaWaK'01). Munich, Germany, September 2001.

[Denos ,2004] :N.Denos ,C.Berrut ,L. Gallardo-Lopez ,A. Nguyen . « COCoFil : Une plateforme de filtrage collaboratif orientée vers la communauté », Actes de la 1ère Conférence en Recherche d'Information et Applications (CORIA'04), Toulouse, France, p. 9-26 ,2004.

[DO Minh Chau, 2007] :Mémoire de fin d'étude . « Vers une approche personnalisée de la recherche d'informations »

[Freund, 1996] : Freund, Y. and Schapire, R. (1996). Experiments with a new boosting algorithm" .In Machine Learning : Proceedings of the Thirteenth International Conference, pages 148–156.

[Jin,2003] : Jin X., Mobasher B., "Using semantic similarity to enhance item-based collaborative filtering", DePaul University, 2003.

[Hamid, 2004] :T.Hamid. « Formalisation et spécification d'un système de filtrage incrémental d'information » ,2004.

[Han et Kamber, 2001] : Han, J. et Kamber, M. (2001). Data Mining : Concepts and Techniques. Morgan Kaufmann, San Francisco, California, USA.

[Herlocker ,1999]:J.Herlocker ,A. Konstan ,A. Borchers ,J. Riedl , « An Algorithmic Framework for Performing Collaborative Filtering », Proceedings of the 22nd International ACM Conference on Research and Development in Information Retrieval (SIGIR'99), USA, p. 230-237,1999.

[Herlocker ,2000] :J.Herlocker . « Understanding and Improving Automated Collaborative Filtering Systems », Ph.D Dissertation, University of Minnesota, 2000.

[Herlocker et al,2000]:J.Herlocker ,J. Konstan ,J. Riedl . « Explaining Collaborative Filtering Recommendations », Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW'00), Pennsylvania, USA, p. 241-250,2000.

[Houda,2009] :O.Houda. « Amélioration du processus de démarrage à froid dans les systèmes de filtrage d'information collaboratif » . 2009

[Goldberg ,1992] :D.Goldberg ,B. Oki , D.Nichols ,D. Terry . « Using Collaborative Filtering to Weave an Information Tapestry », Communications of the ACM, vol. 35 (12), p. 61-70,1992.

[Ilham, 2010]: Mémoire de fin d'étude. Vers une approche comportementale de recommandation : apport de l'analyse des usages dans un processus de personnalisation.2010

[Good ,1999] : N.Good , J.Schafer ,J.Konstan ,A. Borchers ,B. Sarwar ,J. Herlocker ,J.Riedl J.Combining . « collaborative filtering with personal agents for better recommendations »,

Proceedings of the 16th National Conference on Artificial Intelligence, Orlando, USA, p. 439-446,1999.

[Gruber ,1993] :Gruber T.R. (1993). "Towards Principles for the Design of Ontologies used for Knowledge Sharing". In N. Guarino (Ed.). Int. Workshop on Formal Ontology, Padova, Italy.

[Kohrs,2001] : A.Kohrs , B.Merialdo. « Improving Collaborative Filtering for New Users by Smart Object Selection », Proceedings of International Conference on Media Features (ICMF), Italy, 2001.

[Krulwich et Burkey, 1996] :Krulwich, B. et Burkey, C. (1996). Learning user information interests through extraction of semantically significant phrases. In Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access. Stanford, CA.

[Lang, 1995]: Lang, K. (1995). Newsweeder : Learning to filter netnews. In Proceedings of the 12th International Conference on Machine Learning (ICML95), pages 331–339.

[Latifa, 2010] : Latifa Baba-Hamed, Réda Soltani et Kamel Sabri . « Construction d'une ontologie pour la recommandation de films à un utilisateur » .IC 2010: Atelier GBPOnto 08 juin 2010

[Laublet, 2004] : LAUBLET. « Introduction au Web sémantique ». Rennes : URFIST, 2004. Support de formation (sous Power Point) pour un stage URFIST, 26 mai 2004.

[Lin ,1998] : Lin D., "An Information-Theoretic Definition of Similarity", In Proceedings of the 15th International Conference on Machine Learning,

[Linden,2003] :G.Linden, S.Brent, J.York.«Amazon.com recommendations:Item-to-item collaborative filtering», IEEE internet computing, vol. 7, n°1, p. 76-80, 2003.

[Lumineau,2002] : N. 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.

[Lumineau, 2003] : Nicolas Lumineau, "Un tour d'horizon du filtrage collaboratif", 2003

[Malone et al,1987] :T.Malone, K.Grant, F.Turbak, S.Brobst, et M.Cohen. « Intelligent information sharing systems », Communications of the ACM, 30(5) :390–402, 1987.

[Maltz,1995] :D.Maltz ,E.Ehrlich . « Pointing the way: Active collaborative filtering », Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'95), USA, 1995, p. 202-209,1995.

[Melville ,2002]: Melville,Raymod J.Mooney et Ramadas Negarajan .2002."Content-boosted collaborative filtring for improved recommendations".In Eighteenth national conference on Artificial intelligence :American for ArtificialvIntelligence.

[Maria, 2005] :G.Maria . « Accès à l'information par un système de filtrage collaboratif contrôlé » , 2005 .

[Maron et Kuhns,1960] : M. Maron and J. Kuhns. « On relevance, probabilistic indexing and information retrieval ». Journal of the ACM, 7(3) :216–244, July 1960.

[Middelton, 2004] :S.Middleton ,H. Alani , C.De Roure . « Ontological User Profiling in Recommender Systems », ACM Trans. Information Systems, vol. 22, no. 1, pp. 54-88, 2004.

[Miller,1997] :D.Miller,J.Maltz, R.Herlocker ,A.Gordan ,A.Riedl,B.Konstan . « GroupLens: applying collaborative filtering to Usenet News », Communications of the ACM, vol. 40, n° 3, p. 77-87, mars 1997.

[Mizzaro,1997] :S.Mizzaro. « Relevance : The whole history. Journal of the AmericanSociety for Infomration Science », 49(9) :810–832, 1997.

[Mobasher,2004] :B.Mobasher, X.Jin, Y.Zhou. «Semantically enhanced collaborative filtering on the Web», Book chapter, Web Mining: FromWeb to SemanticWeb, 2004.

[Mooney,1998]:R.Mooney ,P.Bennett ,L.Roy . « Book Recommending Using Text Categorization with Extracted Information », Proc.Recommender Systems Papers from 1998 Workshop, Technical Report, WS-98-08,1998.

[MovieLens] : MovieLens,http://movielens.umn.edu, http://www.grouplens.org.

[Nguyen,1998] :H.Nguyen , P.Haddawy . « The Decision-Theoretic Video Advisor, Working Notes of the AAAI-98 Workshop on Recommender Systems », Wisconsin, USA, 1998, p. 77-80, 1998.

[Park, 2006]: Park S., Pennock D., Madani O., Good N., DeCoste D., «Naïve Filterbots for Robust Cold-Start Recommendations», Proc. Of KDD'06, USA, 2006.

[Passin, 2004]: Thomas B. Passin, "Explorer's Guide to the Semantic Web", 2004.

[Philip , 1999]:Philip Chan. A non-invasive learning approach to building Web user profiles. In Workshop on Web usage analysis and user profiling, Fifth International Conference on Knowledge Discovery and Data Mining, August 1999.

[Prem, 2010] : Prem Melville,Vikas Sindhwani."Machine learning"IBM T.J.Watson Research center.

[Resnik ,1994] :P. Resnick ,N. Iacovou , M.Suchak , P.Bergstrom ,J. Riedl , GroupLens. « An Open Architecture for Collaborative Filtering of Netnews », Proceedings of the Conference on Computer Supported Cooperative Work (CSCW'94), NC, USA, 1994.

[Resnik ,1994] : Resnik P., "Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language", Journal of Artificial Intelligence Research JAIR, ISSN 11076 – 9757, Volume 11, pp. 95-130, 1999.

[Salton,1971] :G. Salton. « The SMART Retrieval System - Experiment in Automatic Document Processing ». Englewood Cliffs, NJ : Prentice-Hall, 1971.

[Salton,1983] :G.Salton. « Introduction to modern information retrieval ». New York, McGraw-Hill, 1983.

[Samia, 2007] :B.Samia . « Modélisation hybride du profil utilisateur pour un système de filtrage »,2007.

[Sarwar et al., 2000b] :Sarwar, B. ; Karypis, G. ; Konstan, J. et Riedl, J. (2000b). "Application of dimensionality reduction in recommender system - a case study". In ACM WebKDD 2000 Web Mining for ECommerce Workshop.

[Schein,2001] :A.Schein ,A. Popescul ,L. Ungar . « Generative Models for Cold-Start Recommendations », Proceedings of the 2001 SIGIR Workshop on Recommender Systems, USA, 2001.

[Shardanand et Maes,1995] : Shardanand, U. et Maes, P. (1995). Social information filtering: algorithms for automating "word of mouth". In Proceedings of the SIGCHI

conference on Human factors in computing systems (CHI'95), pages 210–217, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co.

[Su et Khoshgoftaar, 2009]: Su, X. et Khoshgoftaar, T. (2009). A survey of collaborative filtering techniques. Advances in Artificial Intelligence, Janvier 2009:1–20.

[Tufféry, 2007] :Tufféry, S. (2007). Data mining et statistique décisionnelle : l'intelligence des données. Editions Ophrys.

[Ungar ,1998] :L.Ungar ,et D. Foster. « Clustering methods for collaborative filtering », 1998

[Viappiani et al., 2006] : Viappiani, P. ; Faltings, B. et Pu, P. (2006). Preference-based search using example-critiquing with suggestions. Journal of artificial intelligence Research,

27:465–503.

[Wang, 2006]: Jun Wang, Marcel J.T. Reinders (Delft University of Technology), Arjen P. de Vries (CWI), "Unifying User-based and Item-based Collaborative Filtering Approaches by Similarity Fusion", 2006.

[Ziegler, 2004] :C.Ziegler , L. Schmidt- Thieme, G .Lausen. «Exploiting semantic product descriptions for recommender systems», ACM SIGIR Semantic and Information Retrieval Workshop, 2004.

[Ziegler, 2007] :C.Ziegler , J. Golbeck. «Investigating interactions of trust and interest similarity», Decision Support Systems, vol. 43, n° 2, p. 460-475, 2007.

[Zuber, 2006] :V.Zuber, B.Faltings. «Inferring User's Preferences using Ontologies», p. 1413-1418,2006.