# Modèle d'espaces de communautés orienté vers la diversité de recommandations pour les systèmes de filtrage

An-Te Nguyen, Nathalie Denos, Catherine Berrut

# Modèle des espaces de communautés orienté vers la diversité des recommandations pour les systèmes de filtrage

An-Te Nguyen*, Nathalie Denos* et Catherine Berrut*

* Laboratoire CLIPS-IMAG
{an-te.nguyen, nathalie.denos, catherine.berrut}@imag.fr

***Résumé***

*Les systèmes de filtrage ont pour but de distribuer des informations de façon personnalisée aux utilisateurs, tout en s'adaptant en permanence au besoin en information de chacun. Dans un système de filtrage hybride s'appuyant sur le filtrage collaboratif, la production de recommandations se base sur des communautés d'utilisateurs qui sont généralement formées conformément au seul critère de proximité des évaluations des utilisateurs sur les recommandations reçues dans le passé. De plus ces communautés restent généralement implicites.*

*Nous proposons le modèle des espaces de communautés multicritères et explicites, et des mesures se basant sur la théorie des ensembles d'approximation pour analyser la dépendance entre les critères de formation des communautés. Notre modèle des espaces de communautés permet de diversifier les recommandations qui peuvent émaner de communautés variées. Les mesures permettent de comparer les critères entre eux afin de déterminer un ordre de priorité sur les critères dans la tâche d'amélioration du positionnement des utilisateurs dans les communautés.*

***Mots-clés*** *: Système de filtrage hybride, espace de communautés, vecteur de positionnement, théorie des ensembles d'approximation.*

***Abstract***

*Recommender systems intend to provide relevant information based on user preference. Most of them rely on collaborative filtering that compares users on the basis of their ratings in order to group them into communities and produce recommendations with respect to these communities. We propose a model that allows for the explicit management of multi criteria community spaces, in order to diversify recommendations as they arise from communities based on different criteria. We also define measures based on the Rough Sets Theory, to compare criteria in order to determine an order to follow when trying to improve the position of users in community spaces.*

***Keywords****: Hybrid recommender system, community space, position vector, Rough Sets Theory.*

# 1. INTRODUCTION

## 1.1. Système de filtrage

Aujourd'hui les systèmes de filtrage adaptatif, ou systèmes de recommandation, sont utilisés dans divers domaines comme la recherche documentaire, le commerce électronique, les loisirs, etc. Leur objectif est de filtrer un flux entrant d'informations (documents) de façon personnalisée pour chaque utilisateur, tout en s'adaptant en permanence au besoin d'information de chacun [14]. Pour cela, les moteurs de systèmes de recommandation gèrent des profils d'utilisateurs permettant de choisir quels documents transmettre à chacun. Au cours du temps, les profils des utilisateurs sont mis à jours sur la base du retour de pertinence que les utilisateurs fournissent sur les documents reçus.

Il existe trois grandes approches de filtrage : basé sur le contenu, collaboratif et hybride. Le filtrage basé sur le contenu compare les nouveaux documents au profil de chaque utilisateur et recommande ceux qui sont le plus proche [14]. Le filtrage collaboratif compare les utilisateurs entre eux sur la base de leurs jugements passés pour créer des communautés, et chaque utilisateur reçoit les documents jugés pertinents par sa communauté [2]. Le filtrage hybride combine filtrage basé sur le contenu et filtrage collaboratif pour exploiter au mieux les avantages de chacun [3] : en général le système hybride gère des profils d'utilisateurs orientés contenu, et la comparaison entre ces profils donne lieu à la

formation de communautés d'utilisateurs permettant le filtrage collaboratif.

## 1.2. Problématique

Dans un système de filtrage hybride, la gestion des communautés joue un rôle clé dans la production de recommandations puisque la qualité des recommandations collaboratives dépend fondamentalement de la qualité des communautés formées par le système. Ainsi, nous abordons dans cette section deux questions importantes relatives à cette tâche de gestion : la formation des communautés et le positionnement des utilisateurs au sein des communautés.

**Formation des communautés selon un unique critère**. Dans la plupart des systèmes de filtrage[1], les communautés sont généralement formées selon un seul critère [22], par exemple la proximité des évaluations des utilisateurs. Pourtant il existe de multiples critères sur lesquels la formation des communautés peut s'appuyer.

Prenons l'exemple d'un système de recommandation de films dans lequel le profil d'un utilisateur [1] contient sa profession, la ville où il habite, son genre de film préféré et ses évaluations sur certains films. Pour chacun de ces attributs, on peut regrouper les utilisateurs selon leur proximité relativement à cet attribut ou critère. Autrement dit, le système peut construire une partition des utilisateurs, ou *espace de communautés*, pour chacun des critères Profession, Ville et Genre préféré en plus de l'espace de communautés créé par le critère classique Evaluation basé sur les évaluations d'utilisateurs. Ainsi, l'utilisateur ayant pour valeur « Chercheur » pour le critère Profession, « Paris » pour le critère Ville et « Documentaire » pour le critère Genre préféré et une liste de ses évaluations passées, appartient simultanément à quatre communautés différentes selon ces quatre critères.

La *table de communautés* illustrée dans le tableau 1 est construite à partir de ces quatre critères, dont chaque valeur est une étiquette de communauté. Chaque colonne de la table correspond à un espace de communautés relatif à un critère, et chaque ligne contient les communautés auxquelles appartient un utilisateur particulier. La figure 1 nous montre que les utilisateurs sont associés très différemment les uns

[1] Dans le reste de l'article, le terme « système de filtrage » désigne, sauf précision autre, un système hybride qui combine le filtrage collaboratif avec d'autres techniques de filtrage surtout avec le filtrage basé sur le contenu.

aux autres selon le critère choisi. L'utilisateur $u_3$ est proche de $u_4$ pour le critère Profession mais éloigné pour le critère Genre préféré.

| Critère (espace) / Utilisateur | Profession | Ville | Genre préféré | Evaluation |
|---|---|---|---|---|
| $u_1$ | Commerçant | Paris | Aventure | Groupe 1 |
| $u_2$ | Chercheur | Paris | Aventure | Groupe 4 |
| $u_3$ | Chercheur | Paris | Documentaire | Groupe 2 |
| $u_4$ | Chercheur | Paris | Policier | Groupe 1 |
| $u_5$ | Chercheur | Paris | Policier | Groupe 4 |
| $u_6$ | Chercheur | Paris | Policier | Groupe 3 |
| $u_7$ | Chercheur | Paris | Fiction | Groupe 5 |
| $u_8$ | Chercheur | New York | Documentaire | Groupe 5 |
| $u_9$ | Commerçant | New York | Documentaire | Groupe 5 |
| $u_{10}$ | Commerçant | Londres | Documentaire | Groupe 3 |
| $u_{11}$ | Commerçant | Londres | Documentaire | Groupe 2 |
| $u_{12}$ | Commerçant | Londres | Documentaire | Groupe 3 |

**TAB. 1 –** *Table de communautés d'un système de recommandation de films*

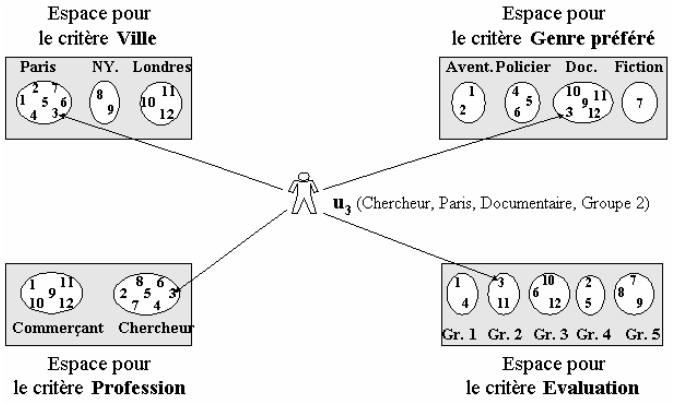


**FIG. 1 –** *Espaces de communautés pour chaque critère*

Dans ce contexte de multiplicité de critères, la question se pose du choix du critère à utiliser pour former les communautés. Nous avons montré dans [16] que les espaces de communautés formés sur les critères Genre préféré et Evaluation sont différents, d'où l'idée d'exploiter cette différence pour enrichir la production de recommandations. Ainsi, un

utilisateur peut recevoir les recommandations de chacune des communautés auxquelles il appartient. Dans notre exemple, l'utilisateur $u_3$ peut recevoir les films envoyés via les communautés Chercheur, Paris et Documentaire outre les films émanant de la communauté Groupe 2 basée sur le critère Evaluation. Cette approche conduit à des recommandations plus diversifiées qui peuvent être exploitées de façon sélective dans une stratégie adaptée à la situation des utilisateurs. Convaincus de l'intérêt de cette approche, nous proposons dans cet article une formalisation adéquate de ces espaces de communautés.

**Positionnement difficile d'utilisateurs dans les espaces de communautés**. En général, la qualité du positionnement des utilisateurs dans les espaces de communautés dépend fondamentalement de la qualité des valeurs données pour chaque utilisateur à chaque critère. Certains critères demandent beaucoup d'efforts de la part des utilisateurs, et peuvent être coûteux également pour le système [14]. Par exemple, un nouvel utilisateur peine à définir son genre de film préféré, qui est par ailleurs susceptible d'évoluer. De même, évaluer un grand nombre de films est une tâche lourde pour l'utilisateur [21], [14], et c'est pourtant ainsi que les communautés sont formées pour le critère Evaluation.

Cela conduit à l'absence de valeur pour un ou plusieurs critères (par exemple on ne connaît pas le genre préféré) et l'existence de valeurs douteuses ou périmées (par exemple les évaluations sont en très petit nombre), d'où des difficultés dans le positionnement des utilisateurs dans les espaces de communautés [17]. Nous apportons ici une contribution pour résoudre ce second problème, en définissant des bases utiles à l'élaboration d'une stratégie d'amélioration de ces positionnements.

## 1.3. Objectifs et plan de l'article

L'objectif de notre article est de répondre aux deux problèmes évoqués dans la section précédente, qui concernent la formation des communautés et le positionnement des utilisateurs dans les espaces de communautés.

**Formation multicritère de communautés**. Dans cet article nous formalisons la formation des « espaces de communautés » selon des critères variés : les communautés ne sont plus seulement formées sur la base des jugements d'utilisateurs mais aussi sur tous les autres critères de rapprochement entre utilisateurs qui sont disponibles dans le système. Par conséquent, un utilisateur appartient à une communauté dans chaque espace (voir Figure 1), ce qui est traduit par son « vecteur de positionnement ».

**Positionnement des utilisateurs dans les espaces de communautés**. Nous étudions également la possibilité de rattacher un utilisateur à une communauté dans un espace relatif à un certain critère à partir de ses communautés déjà connues dans d'autres espaces. Nous définissons des moyens pour analyser les relations existant entre les critères. Ce sont des mesures permettant de comparer les divers critères selon : a) leur capacité à se déterminer les uns les autres (par exemple, Ville = « New York » et Genre = « Documentaire » déterminent Evaluation = « Groupe 5 ») ; et b) lesquels sont prioritaires dans le système (par exemple, on connaît toujours Profession et Ville mais pas toujours Genre et Evaluation). Nous proposons d'ordonner ces critères dans le but d'élaborer des stratégies pour améliorer le positionnement de l'utilisateur dans les divers espaces de communautés, et à terme améliorer les recommandations.

Nous nous appuyons sur la théorie des ensembles d'approximation (« rough sets ») [16], [17], [18] pour formaliser les espaces de communautés. Cette théorie nous donne aussi des moyens efficaces pour analyser la dépendance entre un critère clé, ou décision, a priori fixé et les critères restants. Nous proposons une extension de cette théorie pour définir les mesures destinées à comparer les critères sans fixer aucune décision a priori. Enfin nous présentons la mise en œuvre et le test de notre modèle sur le jeu de données réelles MovieLens [15].

## 2. CONTEXTE ET ETAT DE L'ART

Dans cette partie, nous décrivons les principales études sur la notion de communauté dans les systèmes de recommandation et nous présentons brièvement les notions de base de la théorie des ensembles d'approximation sur laquelle s'appuie notre modèle des espaces de communautés.

### 2.1. Notion de communauté

Dans le contexte d'un système de filtrage, la notion de communauté est définie comme un ensemble d'utilisateurs qui sont proches les uns des autres selon un critère donné de comparaison. Dans les systèmes de recommandation existants, les communautés sont en général formées par la proximité des évaluations, explicites ou implicites, des utilisateurs. Selon le principe du filtrage collaboratif, ces communautés sont utilisées pour produire des recommandations [2], [14], [22].

Actuellement, l'approche basée mémoire est la plus populaire pour la formation des communautés [2]. Dans cette approche, le système traite la base de données entière des évaluations pour mesurer la proximité entre utilisateurs, et on utilise souvent la corrélation de Pearson dans le calcul (voir [2] pour une description de cette corrélation en contexte du filtrage collaboratif). Ces techniques basées mémoires sont simples à implémenter, et efficaces dans la plupart des cas. Néanmoins, elles souffrent du problème du coût de calcul puisque le système doit traiter à chaque fois la base de données entière. De plus, ces techniques nécessitent un grand nombre de documents jugés en commun entre deux utilisateurs afin d'obtenir des résultats fiables.

Pour créer des communautés, on trouve également l'approche basée modèle [2]. D'abord, le système essaie de construire par apprentissage un modèle probabiliste à partir des évaluations d'utilisateurs. Ensuite, le système applique ce modèle pour prédire la communauté la plus probable pour un nouvel utilisateur en se basant sur ses évaluations. Par exemple, on peut utiliser la classification Bayésienne pour la prédiction des communautés. L'avantage de cette approche est la compacité du modèle, et donc la rapidité du calcul de prédiction. En revanche, ces techniques sont compliquées et le processus d'apprentissage est long.

Dans [22], Perugini et al. donnent un point de vue social sur les systèmes de recommandation, sous l'angle de l'effort à produire pour établir des relations entre utilisateurs. Par exemple, dans l'approche de fouille et d'exploration de structure (*Mining and Exploiting Structure*) [13], le système transforme un réseau biparti $R$ (voir Figure 2), qui représente une matrice d'évaluations ayant deux classes de nœuds {personne $p_i$} et {document $d_j$}, en un réseau social uniparti $G_s$. Enfin, le système rattache les deux réseaux $R$ et $G_s$ en $G_R$ pour l'exploration et l'exploitation dans la production de recommandations. Dans cet article, Mirza et al. proposent la technique « hammock jump » qui relie deux personnes ayant un certain nombre d'évaluations communes pour induire le réseau social $G_s$.

Cette approche permet d'une part de reconnaître et d'explorer des structures dans l'ensemble des évaluations, et d'autre part de former des communautés par transitivité sans avoir besoin d'évaluations en commun entre les utilisateurs.

Il ressort des travaux existants sur les systèmes de recommandation, que les communautés sont formées sur un unique critère, qu'elles demeurent implicites, et qu'elles ne sont pas exploitées en tant que telles. Pourtant, on trouve généralement dans les systèmes une variété de

critères sur lesquels appuyer la formation de communautés : données démographiques, domaines d'intérêt, jugements qualitatifs émis sur des documents, préférence de livraison et de sécurité, etc. [1]. En utilisant chacun de ces critères comme critère de formation, le système peut créer autant d'espaces de communautés (voir Figure 1), et permettre ainsi à un utilisateur d'appartenir à autant de communautés qu'il y a de critères pour les former.

Le fait que les communautés restent souvent implicites et ne sont considérées que comme des résultats intermédiaires dans la production de recommandations, limite les explications sur les recommandations que le système envoie aux utilisateurs. Les études expérimentales de Herlocker et al. [7] ont montré que l'explication des recommandations améliore significativement la confiance et la motivation des utilisateurs à fournir des évaluations, ce sans quoi le système ne peut pas former de meilleures communautés.

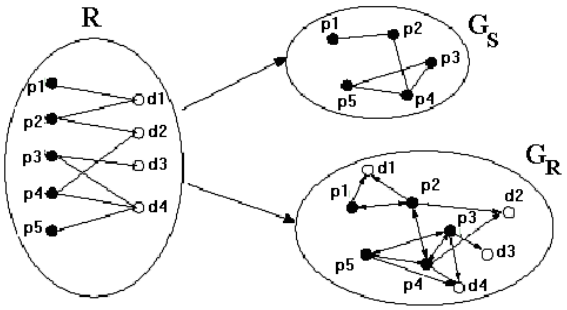


**FIG. 2** – *Fouille et exploitation de structure*

## 2.2. Espaces de communautés et théorie des ensembles d'approximation

Notre problématique requiert le traitement de données symboliques comme les étiquettes dans la table de communautés (voir Tableau 1), ainsi que la prise en compte et l'analyse des imperfections de ces données. En effet, les situations problématiques sont souvent liées à l'absence de certaines informations relatives à l'utilisateur : par exemple un nouvel utilisateur est confronté au démarrage à froid car il n'a pas

encore évalué de document, et/ou ne sait pas définir son besoin en termes de contenu.

Plus généralement, dans ce contexte des données multidimensionnelles, nous devons prendre en compte la dépendance entre les dimensions, ou critères. Cette problématique est étudiée dans plusieurs domaines tels que les bases de donnés [11], OLAP (*Online Analytical Processing*) [1], MDS (*Multidimensional Scaling*) [4] et la fouille de données [7].

Après une étude des théories applicables aux problèmes de données manquantes [5], [12], nous avons choisi la théorie des ensembles d'approximation [16], [17], [18], [23] car elle permet de traiter des données nominales et ne requiert aucune information a priori. Cette théorie permet aussi d'identifier et exploiter les dépendances entre les attributs des données. Ici, les objets étudiés sont les utilisateurs, leurs attributs sont les divers critères de formation des espaces de communautés, et les données sont les étiquettes des communautés auxquelles ils appartiennent. Dans le cadre de cette théorie, nous souhaitons compenser l'absence de certaines valeurs d'attributs dans le positionnement d'utilisateurs en exploitant au mieux les attributs où les valeurs sont disponibles. Par exemple connaissant Ville, Profession et Genre, nous aimerions pouvoir proposer une communauté dans l'espace relatif au critère Evaluation sans même demander d'évaluations à l'utilisateur.

Dans la suite, nous présentons les notions de base de la théorie des ensembles d'approximation qui sont utiles pour notre modèle (voir les références ci-dessus pour des informations plus complètes).

### 2.2.1. Représentation des données et règle de décision

Une table d'information $T$ est caractérisée par deux ensembles non vides $<U, A>$, où $U$ est l'univers des objets et $A$ contient des attributs (critères) (voir Tableau 1). L'ensemble des attributs $A$ est divisé en deux : $D$ contenant un seul attribut dit « décision » et $C$ contenant les attributs restants dits « conditions ». La table $T$ est dite « table de décision », et chaque ligne de la table est considérée comme une règle de décision donnée sous la forme : $C \rightarrow D$

Par exemple, on a dans le tableau 2, où Evaluation est pris comme attribut de décision, la règle $u_7$ qui dit : « Un chercheur parisien qui aime

les films de genre Fiction appartient à la communauté libellée Groupe 5 dans l'espace Evaluation ».

Soit $P \subseteq A$ et $R_P$ la relation d'équivalence sur $U$ où les objets sont en relation s'ils ont les mêmes valeurs pour tous les attributs de $P$ :

$$\forall u, u' \in U, \quad u \, R_P \, u' \;\Leftrightarrow\; (\forall a \in P, T[u, a] = T[u', a])$$

La classe d'équivalence de u suivant $R_P$ est notée $[u]_P$, et l'ensemble quotient $U/R_P$ comporte toutes les classes d'équivalence de $U$ suivant $R_P$ (partition de $U$ selon les valeurs prises sur $P$). Par exemple, pour $P$ = {Ville, Genre préféré}, $U/R_P$ contient les six classes d'équivalence suivantes (voir Tableau 2) :

$[u_1]_P = \{u_1, u_2\}$

$[u_3]_P = \{u_3\}$

$[u_4]_P = \{u_4, u_5, u_6\}$

$[u_7]_P = \{u_7\}$

$[u_8]_P = \{u_8, u_9\}$

$[u_{10}]_P = \{u_{10}, u_{11}, u_{12}\}$

Pour l'attribut de décision $D$, la classe d'équivalence $[u]_D$ est appelée « concept ». Dans l'exemple précédent, avec $D$ = {Evaluation}, on a cinq concepts :

$X_1 = \{u_1, u_4\}$

$X_2 = \{u_3, u_{11}\}$

$X_3 = \{u_6, u_{10}, u_{12}\}$

$X_4 = \{u_2, u_5\}$

$X_5 = \{u_7, u_8, u_9\}$

Supposons que $P$ = {Ville, Genre} et $D$ = {Evaluation}. On remarque que les règles $u_1$ et $u_2$ sont contradictoires puisqu'elles ont la même prémisse : Ville = « Paris » et Genre = « Aventure » mais qu'elles donnent deux concepts différents : « Groupe 1 » et « Groupe 4 ».

On définit la notion de « règle certaine » $u$ comme une règle dont la classe d'équivalence par $P$ est incluse dans celle par $D$ :

$$[u]_P \subseteq [u]_D \quad (1)$$

Ce sont des règles pour lesquelles un utilisateur est certainement positionné dans une communauté relative à $D$ en se basant sur $P$. La « région positive » de $U/R_P$, notée $POS_P(D)$, est l'union de toutes les règles certaines de la table $T$. Dans l'exemple ci-dessus, $u_3$, $u_7$, $u_8$ et $u_9$ sont des règles certaines puisque :

$[u_3]_P \subseteq [u_3]_D$

$[u_7]_P \subseteq [u_7]_D$

$[u_8]_P \subseteq [u_8]_D$

$[u_9]_P \subseteq [u_9]_D$

Les règles certaines sont donc celles qui donnent toujours la même valeur pour le critère Evaluation étant donné une valeur pour Ville et pour Genre (voir Tableau 2).

| U | Profession | Ville | Genre | Evaluation | $POS_p$(Eval.) |
|---|---|---|---|---|---|
| $u_1$ | Commerçant | Paris | Aventure | Groupe 1 | |
| $u_2$ | Chercheur | Paris | Aventure | Groupe 4 | |
| $u_3$ | Chercheur | Paris | Documentaire | Groupe 2 | x (inclus) |
| $u_4$ | Chercheur | Paris | Policier | Groupe 1 | |
| $u_5$ | Chercheur | Paris | Policier | Groupe 4 | |
| $u_6$ | Chercheur | Paris | Policier | Groupe 3 | |
| $u_7$ | Chercheur | Paris | Fiction | Groupe 5 | x |
| $u_8$ | Chercheur | New York | Documentaire | Groupe 5 | x |
| $u_9$ | Commerçant | New York | Documentaire | Groupe 5 | x |
| $u_{10}$ | Commerçant | Londres | Documentaire | Groupe 3 | |
| $u_{11}$ | Commerçant | Londres | Documentaire | Groupe 2 | |
| $u_{12}$ | Commerçant | Londres | Documentaire | Groupe 3 | |

$X_5$

**TAB. 2 –** *Exemple de région positive (D = {Evaluation} et P = {Ville, Genre})*

### 2.2.2. Dépendance d'un attribut de décision, consistance de la table et signification d'attributs de condition

La « dépendance » de l'attribut de décision $D$ par rapport à l'ensemble $P \subseteq C$ peut être mesurée par :

$$\gamma(P, D) = \frac{|POS_P(D)|}{|U|} \quad (2)$$

Ce coefficient traduit la part des objets donnant lieu à des règles certaines. Pour l'ensemble des attributs de condition $C$, $\gamma(C, D)$ exprime la « consistance » de la table $T$ vis-à-vis de l'attribut de décision $D$ par les attributs de condition $C$.

La théorie des ensembles d'approximation cherche à réduire la taille de la prémisse dans les règles en tenant compte de la performance de l'induction par règles. On définit ainsi la « réduction » de $C$ comme le sous-ensemble $P$ de $C$ comprenant l'ensemble minimal d'attributs de condition permettant de préserver la région positive : $POS_P(D) = POS_C(D)$

Une même table de décision peut donner plusieurs réductions. Théoriquement, les réductions peuvent être calculées en utilisant la matrice d'indiscernabilité [25] ; la complexité de ce calcul est en $O(m.n^2)$ où m est la taille de l'ensemble U et n est le nombre d'attributs. Il existe dans la littérature des méthodes heuristiques de calcul des réductions [15] qui rendent réaliste la théorie des ensembles d'approximation.

Par ailleurs, on peut mesurer l'importance d'un sous-ensemble d'attributs de condition $P \subset C$ sur l'intervalle [0,1] par sa « signification » qui mesure l'impact de sa suppression sur la consistance de la table $T$ :

$$\sigma_{C,D}(P) = 1 - \frac{\gamma(C \setminus P, D)}{\gamma(C, D)} \quad (3)$$

Si $P$ est une réduction de $C$, on a $\sigma_{C,D}(P) = 1$, aussi noté $\sigma(P)$ si $C$ et $D$ sont fixés.

## 3. MODELE DES ESPACES DE COMMUNAUTES

Nous présentons ici notre modèle de gestion des communautés dans un système de filtrage, basé sur la théorie des ensembles d'approximation. En général, les valeurs dans la table des communautés $T$ (voir Tableau 1) sont obtenues à partir des données disponibles dans les profils des utilisateurs. Certains critères sont obtenus de façon directe (exemple :

« Ville ») ; pour certains autres on doit regrouper plusieurs valeurs (exemple : « Tranche d'âge ») pour former des communautés plus significatives ; pour d'autres encore on doit faire appel à une méthode de classification [10] en utilisant une mesure de similarité (exemple : « Evaluation »). Nous avons montré dans [16] comment former ces communautés afin de remplir cette table.

Nous montrons ici comment formaliser la notion d'espace de communautés, puis celle de vecteur de positionnement. Ensuite, nous abordons l'induction de communautés par règles. Enfin nous définissons les mesures destinées à comparer entre eux les critères pour l'induction de communautés.

## 3.1. Espace de communautés et vecteur de positionnement

Pour tout critère $a \in A$, on note $V_a$ l'ensemble des valeurs que l'on trouve dans la table $T = <U, A>$. La partition de $U$ pour le critère $a$, notée $\Omega_a$, est appelée « espace de communautés ». Elle correspond à l'ensemble quotient $U/R_a$ et est composée des groupes d'utilisateurs $Ga_k$ qui prennent la même valeur $v_k$ pour le critère $a$. Les groupes $Ga_k$ correspondent à la notion de communauté, et il existe une bijection entre $V_a$ et $\Omega_a$, car à chaque valeur $v_k$ dans $V_a$ est associée une et une seule communauté $Ga_k$.

En généralisant cette formalisation, on définit la notion de « critère composé » comme un ensemble d'au moins deux critères $P \subseteq A$, et l'« espace de communautés composé » $\Omega_P$ est l'ensemble quotient $U/R_P$. $\Omega_P$ est donc constitué des groupes d'utilisateurs ayant les mêmes valeurs pour tous les critères de $P$ (voir le critère composé $P$ = {Ville, Genre} dans 2.2.1).

On appelle « vecteur de positionnement » de l'utilisateur $u$ la liste des étiquettes de ses propres communautés selon chacun des critères. Ce vecteur correspond à la ligne $T[u]$ et est noté $P_u$. Dans notre modèle, ces vecteurs traduisent le polymorphisme du positionnement des utilisateurs au sein des communautés (voir Figure 1) : un même utilisateur n'aura pas les mêmes « voisins » selon le critère considéré.

Avec cette représentation formelle des espaces de communautés, nous répondons au premier problème identifié en 1.2, en permettant une formation relative à plusieurs critères. Nous illustrons maintenant la façon d'exploiter la théorie des ensembles d'approximation en vue d'améliorer

le positionnement des utilisateurs qui est souvent imparfait (valeurs manquantes ou douteuses dans les vecteurs de positionnement).

## 3.2. Induction de communautés

Pour certains critères, la classification des utilisateurs se réalise facilement par la comparaison directe des valeurs. Par exemple, le système peut diviser les utilisateurs en deux classes selon le sexe masculin ou féminin. Parfois, le système doit appliquer une combinaison des valeurs pour regrouper les utilisateurs, par exemple pour former des communautés par tranches d'âge ou par catégories de profession.

La tâche de formation des communautés pour les critères plus complexes comme les centres d'intérêt ou les évaluations passées devient difficile puisque le système doit procéder par étapes et faire appel à une méthode de classification [16]. Les difficultés dans le positionnement des utilisateurs dans les espaces concernés proviennent souvent de l'imperfection des données dans les profils et/ou de la complexité du calcul. A titre d'exemple, le système de filtrage collaboratif doit souvent faire face au problème du démarrage à froid dont le positionnement d'un nouvel utilisateur dans l'espace du critère Evaluation puisque cet utilisateur n'a encore évalué aucune recommandation [17], [24]. Ce problème s'aggrave encore dans le contexte des communautés multicritères. Ainsi, le vecteur de positionnement d'un utilisateur contient souvent des valeurs (communautés) manquantes ou douteuses.

La performance de la formation des communautés peut être améliorée de façon significative par une méthode permettant d'induire pour un utilisateur la communauté manquante dans un espace à partir de ses communautés déjà connues dans d'autres espaces.

Prenons l'exemple d'un vecteur $P_u$ où il ne manque qu'une valeur. En utilisant la théorie des ensembles d'approximation, on prend comme attribut de décision $D$ le critère correspondant à la valeur manquante. Ensuite, on cherche une règle certaine applicable dans la région positive $POS_{A\backslash D}(D)$ afin d'inférer la valeur pour $D$. Supposons que le nouvel utilisateur $u$ est un chercheur à Paris, qui aime les films documentaires et que sa communauté pour le critère Evaluation n'est pas encore connue car il n'a encore évalué aucun film. Le tableau 2 montre que l'on a la règle certaine $u_3$ :

(Profession = « Chercheur», Ville = « Paris », Genre = « Documentaire ») → (Evaluation = « Groupe 2 »)

Le système peut donc rattacher cet utilisateur à la communauté libellée « Groupe 2 » dans l'espace Evaluation.

En pratique, les règles applicables pour compléter la valeur pour *D* n'existent pas toujours, et on doit alors faire appel à des techniques existantes de remplissage de valeurs manquantes, comme par exemple l'approche probabiliste, techniques que nous ne détaillons pas ici [6].

Il est à noter que dans la théorie des ensembles d'approximation, l'attribut de décision est prédéterminé. Ainsi, cette théorie ne permet en principe d'induire que la valeur du critère choisi à priori comme l'attribut de décision. Au cas où plusieurs valeurs sont manquantes ou douteuses dans un vecteur de positionnement, le système peut corriger toutes ces valeurs en choisissant successivement chacun des critères problématiques comme attribut de décision, et cela dans un ordre arbitraire. Cette méthode naïve n'est pas toujours efficace. Il faut donc élaborer une stratégie pour décider dans quel ordre compléter ou modifier les valeurs, afin de commencer par celles dont le résultat est le plus sûr. Cela revient à déterminer un ordre sur les attributs de décision à prendre en compte successivement. C'est pourquoi nous définissons dans la suite des « mesures de qualité de l'attribut de décision ». Ces mesures permettent de comparer les critères afin de déterminer un ordre selon lequel appliquer la technique de correction souhaitée sur les vecteurs de positionnement (voir Figure 3), partant du principe qu'une table de bonne consistance donnera lieu à de bonnes corrections.

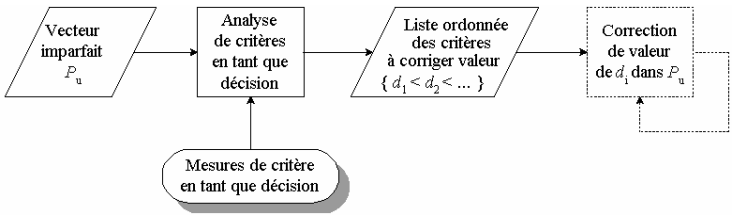


**FIG. 3** – *Mécanisme générique de correction d'un vecteur de positionnement*

## 3.3. Mesures de qualité de l'attribut de décision

Nous proposons trois mesures qui s'appuient sur la consistance de la table et la signification des attributs de condition (voir 2.2.2).

**Mesure basée sur la consistance**. Etant donné l'attribut de décision $D$ et les attributs de condition $C = A \setminus D$, $\gamma(C, D)$ reflète la consistance de la table $T$ par rapport à l'attribut de décision $D$. Notre première mesure de qualité de l'attribut de décision est basée sur cette consistance.

Soit deux attributs de décision $D_1$ et $D_2$. On dit que l'attribut de décision $D_2$ est meilleur que $D_1$ s'il donne lieu à une table de meilleure consistance. Autrement dit, cette mesure favorise l'attribut de décision qui fournit les règles certaines les plus nombreuses.

$$D_1 \lhd D_2 \Leftrightarrow | POS_{A \setminus D1}(D_1) | \leq | POS_{A \setminus D2}(D_2) | \qquad (4)$$

Prenons l'exemple d'un utilisateur dont les valeurs des critères Evaluation et Genre manquent dans son vecteur de positionnement. Alors, le tableau 3 montre que Genre est meilleur en tant qu'attribut de décision du fait de la plus grande taille de sa région positive. Les colonnes $POS_C$(Genre) et $POS_C$(Eval) indiquent les règles certaines incluses dans les régions positives lorsque Genre et Evaluation sont pris comme attribut de décision, respectivement.

| U | Profession | Ville | Genre | Evaluation | $POS_C$(Genre) | $POS_C$(Eval.) |
|---|---|---|---|---|---|---|
| $u_1$ | Commerçant | Paris | Aventure | Groupe 1 | x (inclus) | x |
| $u_2$ | Chercheur | Paris | Aventure | Groupe 4 | | x |
| $u_3$ | Chercheur | Paris | Documentaire | Groupe 2 | x | x |
| $u_4$ | Chercheur | Paris | Policier | Groupe 1 | x | |
| $u_5$ | Chercheur | Paris | Policier | Groupe 4 | | |
| $u_6$ | Chercheur | Paris | Policier | Groupe 3 | x | |
| $u_7$ | Chercheur | Paris | Fiction | Groupe 5 | x | x |
| $u_8$ | Chercheur | New York | Documentaire | Groupe 5 | x | x |
| $u_9$ | Commerçant | New York | Documentaire | Groupe 5 | x | x |
| $u_{10}$ | Commerçant | Londres | Documentaire | Groupe 3 | x | |
| $u_{11}$ | Commerçant | Londres | Documentaire | Groupe 2 | x | |
| $u_{12}$ | Commerçant | Londres | Documentaire | Groupe 3 | x | |

**TAB. 3** – *Comparaison de critères*

Dans les approches de classification par règles, on utilise souvent les facteurs *support* et *confiance* pour mesurer la qualité des règles applicables présentes dans le classificateur. Le support d'une règle est la proportion d'occurrences de cette règle dans l'ensemble des exemples, et la confiance est la proportion d'occurrences de cette règle par rapport aux occurrences de la prémisse [6]. Nous soulignons ici que dans notre

application de la théorie des ensembles d'approximation, les deux facteurs de qualité des règles sont inhérents à la mesure de consistance de la table : la confiance est toujours égale à 1, et le support correspond à la proportion d'occurrences des règles dans la région positive. Notre mesure peut donc s'appliquer même si d'autres techniques de classification par règles sont utilisées pour induire les communautés.

**Mesures basées sur les réductions approximatives**. Dans la théorie des ensembles d'approximation, les réductions jouent un rôle très important dans l'induction par règles. Néanmoins, la définition originale de cette notion est trop forte, et les réductions qui en résultent sont éventuellement proches de l'ensemble $C$. Souvent, dans le cadre d'un système de filtrage, les données fournies par les nouveaux utilisateurs ne permettent de renseigner que peu de critères, comme par exemple seulement leurs informations personnelles. Par conséquent, le système ne peut pas réaliser la tâche d'induction des communautés inconnues si ces critères ne couvrent aucune réduction.

Donc, nous proposons d'autres mesures de qualité de l'attribut de décision définies à partir des réductions approximatives afin de tenir compte de l'adaptation au contexte applicatif : on cherche les attributs de décision permettant de considérer un sous-ensemble $P$ d'attributs de condition aussi réduit que possible tout en conservant une signification supérieure à un seuil $\theta$.

Etant donné l'attribut de décision $D$ et les attributs de condition $C$, on définit d'abord les réductions approximatives $R_D^{(\theta)}$, étant donné le seuil $\theta$ :

$$R_D^{(\theta)} = \{P \subseteq C \mid \sigma(P) \geq \theta\} \quad (5)$$

Ce sont les sous-ensembles de critères « acceptables » et inclus dans $C$. Par exemple, si $\theta$ est égal à 0,8, $R_{Eval}^{(\theta)}$ contient deux réductions approximatives :

$P_1$ = {Ville, Genre}

$P_2$ = {Profession, Ville}

A partir des ensembles $R_D^{(\theta)}$ on peut définir diverses relations d'ordre sur les attributs de décision, soit en fixant un nombre maximum $\alpha$ d'attributs de condition à conserver, soit en fixant un ensemble $C_0$ des attributs de condition jugés utiles dans un contexte donné :

$$D_1 \lhd D_2 \Leftrightarrow |\{P \in R_{D1}{}^{(\theta)} \mid |P| \leq \alpha\}| \leq |\{Q \in R_{D2}{}^{(\theta)} \mid |Q| \leq \alpha\}| \quad (6)$$

$$D_1 \lhd D_2 \Leftrightarrow |\{P \in R_{D1}{}^{(\theta)} \mid C_0 \subseteq P\}| \leq |\{Q \in R_{D2}{}^{(\theta)} \mid C_0 \subseteq Q\}| \quad (7)$$

La formule (6) favorise l'attribut de décision qui donne de petites réductions, ce qui signifie que l'on a besoin de connaître peu de choses sur l'utilisateur pour induire des communautés. Plutôt que la taille des réductions, la formule (7) prend en compte leur contenu, préférant les attributs de condition demandant peu d'effort à produire par l'utilisateur (son code postal, par exemple).

En pratique, la tâche consistant à déterminer toutes les réductions relatives à un attribut de décision $D$, qui est nécessaire pour analyser sa qualité par (6) ou (7), est un problème NP-complet. On peut limiter a priori la taille de $R_D{}^{(\theta)}$ pour diminuer la complexité du calcul, ou calibrer le seuil $\theta$ pour que les réductions dans $R_D{}^{(\theta)}$ contiennent dans la plupart des cas les critères de $C_0$.

**Mesure basée sur la consistance approximative**. Nous proposons finalement une mesure complémentaire permettant de départager les attributs de décision qui ne sont pas départagées par les mesures définies dans (6) ou (7). Elle repose sur la consistance $\mu$ de la table de communautés par rapport aux réductions approximatives.

$$\mu(D) = \frac{1}{\left|R_D{}^{(\theta)}\right|} \sum_{P \in R_D{}^{(\theta)}} \sigma(P) \quad (8)$$

$$D_1 \lhd D_2 \Leftrightarrow \mu(D_1) \leq \mu(D_2) \quad (9)$$

## 3.4. Utilisation des mesures pour corriger un vecteur de positionnement

Supposons que pour un nouvel utilisateur $u$ les communautés des critères Genre et Evaluation dans le vecteur de positionnement ne sont pas encore connues : $P_u$ = (Commerçant, Londres, _, _)

Pour compléter ce vecteur, le système choisit d'abord le critère Genre comme attribut de décision en utilisant la mesure basée sur la consistance (voir Tableau 3). Ensuite, les règles certaines $u_{10}$, $u_{11}$ et $u_{12}$ permettent

d'inférer « Documentaire » comme la valeur du critère Genre. Alors, on obtient : $P_u$ = (Commerçant, Londres, Documentaire, _)

Enfin, le critère Evaluation pourrait être instancié par « Groupe 3 » qui est la valeur dominante dans les trois règles concernées [6].

Il est à noter que le vecteur initial ne serait pas complètement corrigé si on commençait par le critère Evaluation puisque $POS_C$(Eval) ne contient pas les trois règles $u_{10}$, $u_{11}$ et $u_{12}$, et que l'ordre des critères ne dépend que de la situation et pas des cas particuliers des utilisateurs.

L'application de notre modèle pour traiter le vecteur de positionnement d'un nouvel utilisateur est présentée dans [17].

Nous soulignons ici une différence importante entre notre modèle et d'autres applications de la théorie des ensembles d'approximation. Cette théorie fait l'hypothèse que l'attribut de décision est fixé dès la conception du système, et que l'on essaie de sélectionner à partir d'un ensemble d'apprentissage des réductions d'attributs de condition en conservant la qualité de la table de décision. On peut dire que cette théorie focalise notamment sur la qualité des attributs de condition plutôt que sur celle des attributs de décision.

Au contraire, dans notre modèle aucun attribut de décision n'est prédéfini. Au départ, tous les critères sont sur le même plan, et pendant le temps d'exploitation, le système va choisir parmi les critères le meilleur attribut de décision en analysant les données de référence, voire toutes les données disponibles, afin de réaliser une certaine tâche de filtrage d'information dans une situation particulière.

Enfin, nous voyons que le modèle proposé ne vise pas à remplacer les systèmes de filtrage existants, mais offre des moyens supplémentaires prenant en compte un cadre plus large (communautés multicritères) et des possibilités de positionner des utilisateurs se trouvant dans des situations problématiques.

## 4. MISE EN ŒUVRE ET TEST DU MODELE DANS UN SYSTEME DE RECOMMANDATION DE FILMS

Nos travaux présentés par la suite ne visent pas à calibrer des paramètres ou à montrer la performance d'un système particulier s'appuyant sur notre modèle des espaces de communautés, mais à montrer comment utiliser le modèle proposé dans un système de filtrage

existant. Nous montrons que l'on peut construire d'abord la table de communautés pour un système réel comme par exemple MovieLens [15] qui dispose de données variées pour former des communautés selon divers critères. Nous analysons ensuite ces critères en utilisant les mesures proposées plus haut. Bien que le calcul des ensembles d'approximation soit un problème NP-complet, nous n'avons utilisé aucune méthode heuristique [26] en raison du nombre faible de critères dans le jeu de test et du temps de calcul raisonnable (une dizaine de minutes par analyse).

Par ailleurs, la validation proprement dite de notre modèle des espaces de communautés a été montrée dans [17] au travers du cas typique de l'intégration de nouveaux utilisateurs dans le système. Dans cet article, nous avons développé notre méthode d'induction des communautés pour de nouveaux utilisateurs et avons comparé favorablement les résultats de notre approche avec d'autres approches sur ce même jeu de données MovieLens.

## 4.1. Construction de la table de communautés

Le jeu de données réelles MovieLens [15] fourni par le groupe de recherche GroupLens à l'université de Minnesota contient 100 000 évaluations (1 à 5 étoiles) faites par 943 personnes sur 1 682 films de 09/97 à 04/98.

Il faut d'abord construire la table des communautés des 943 vecteurs de positionnement avec six colonnes de critères : Age, Profession, Géographie, Motivation, Contenu et Evaluation. Pour les quatre premiers critères, la création des espaces $\Omega$ est simple, avec un nombre fixe de communautés dans chaque espace $\Omega$ :

- Critère Age : 5 communautés correspondant à 5 tranches d'âge (7 à 73 ans) ;
- Critère Profession : 7 communautés de catégories de professions ;
- Critère Géographie : 44 communautés via les états des Etats-Unis ;
- Critère Motivation : c'est la moyenne mensuelle du nombre d'évaluations depuis l'inscription, traduisant ainsi la tendance des utilisateurs à fournir des évaluations sur les recommandations reçues ; cela conduit à 5 communautés (motivation très faible, faible, moyenne, bonne et excellente).

Le critère Contenu regroupe les utilisateurs partageant les mêmes intérêts quant aux genres de film, alors que le critère Evaluation les

regroupe selon leur façon de juger les films. Pour ces deux critères, la construction de $\Omega_{Contenu}$ et $\Omega_{Evaluation}$ est plus élaborée [16] : nous appliquons la méthode des fourmis artificielles [8] pour placer les utilisateurs dans un espace en *2D*, puis nous remplaçons la méthode des k-moyennes par une classification ascendante hiérarchique [10] en vue d'obtenir un nombre de communautés flexible.

En pratique, les utilisateurs sont souvent bien regroupés dans $\Omega_{Contenu}$, alors que le critère Evaluation conduit à une dispersion des utilisateurs [16], en raison du faible nombre d'objets jugés en commun [2]. Ainsi, dans l'espace $\Omega_{Contenu}$, le nombre de communautés obtenu par la classification hiérarchique est relativement stable (8 communautés) quand on fait varier le seuil d'entropie, alors que celui de $\Omega_{Evaluation}$ varie fortement. Par conséquent, toutes les mesures proposées dans cet article ont été paramétrées par le nombre de communautés dans $\Omega_{Evaluation}$.

## 4.2. Résultats et analyse

Notre objectif est d'établir un ordre entre les critères, pour savoir lequel utiliser comme critère clé afin de prédire au mieux une valeur manquante ou douteuse dans le vecteur de positionnement d'un utilisateur. Les résultats présentés ici montrent que, dans le cas des données de MovieLens, les critères Evaluation et Contenu sont ceux qui obtiennent la priorité la plus faible comme critère clé, alors qu'ils correspondent aux communautés les plus coûteuses à former. Cela indique que l'on peut limiter ce coût en positionnant progressivement les nouveaux utilisateurs dans ces deux espaces via leurs positions dans les autres espaces [17]. Cette approche améliore donc la performance de la formation multiple des communautés dans un système de filtrage.

Dans le tableau 4 consacré à la mesure basée sur la consistance (formule (4)), on constate que le critère Evaluation est le pire des attributs de décision puisque la consistance du résultat est toujours la plus faible, que Géographie et Age sont les meilleurs critères, et que Profession et Motivation viennent ensuite, avant le critère Contenu. On remarque aussi que plus le nombre de communautés de $\Omega_{Evaluation}$ diminue (de 93 à 10), plus la consistance de la table se dégrade excepté celle relative au critère Evaluation qui montre la tendance inverse.

Nous soulignons ici un phénomène intéressant sur les deux critères Géographie et Contenu, et qui montre l'utilité de la mesure dans ce contexte applicatif. En principe, si le critère choisi comme attribut de décision forme un nombre élevé de petites communautés (ce qui peut être

souhaitable, par exemple, pour faciliter la perception des communautés par les utilisateurs qui en sont membres), la taille de la région positive, et donc la consistance de la table de communautés, risque d'être faible. Au contraire, si on prend comme attribut de décision un critère formant un petit nombre de grandes communautés, on a plus de chance d'obtenir une meilleure consistance. Pourtant, notre mesure a montré que selon les données ce n'est pas toujours le cas. En effet, le critère Géographie domine tous les autres, bien qu'il donne lieu à 44 communautés avec une vingtaine de personnes par communauté. A l'inverse, le critère Contenu, qui ne donne lieu qu'à 8 communautés, n'est pas bien classé (5$^{e}$ position). La mesure proposée permet donc d'ordonner les critères en tenant compte des données elles-mêmes via la consistance de la table, ce qui garantit un meilleur choix que si l'on se contente du simple choix heuristique favorisant le plus faible nombre de communautés.

| | Nombre de communautés pour le critère Evaluation ($\|\Omega_{Evaluation}\|$) | | | | | |
|---|---|---|---|---|---|---|
| | 10 | 21 | 31 | 51 | 79 | 93 |
| *Géographie* | 77,62 | 85,68 | 89,82 | 93,64 | 95,55 | 95,97 |
| *Age* | 76,25 | 83,78 | 86,74 | 91,41 | 93,85 | 94,91 |
| *Profession* | 61,40 | 71,79 | 78,90 | 87,27 | 92,26 | 93,21 |
| *Motivation* | 60,13 | 69,99 | 76,67 | 86,21 | 91,20 | 92,90 |
| *Contenu* | 50,27 | 60,45 | 67,76 | 76,88 | 84,94 | 86,74 |
| *Evaluation* | 46,98 | 46,55 | 46,02 | 45,71 | 45,28 | 45,28 |

**TAB. 4** – *Analyse de la consistance de la table de communautés (%)*

Pour la première mesure basée sur les réductions approximatives (formule (6)), le paramètre $\alpha$ peut varier de 1 à 5 selon le nombre de critères que l'on souhaite intégrer dans les attributs de condition. Pour l'expérience nous avons choisi la valeur 1 pour nous placer dans une situation où l'utilisateur interagit avec les communautés : il comprendra mieux un critère simple qu'un critère composé. La valeur 0,7 de $\theta$ a été choisie en se basant sur le tableau 4. Pour la formule (7), on a choisi Géographie pour $C_0$ à titre d'exemple.

Dans les tableaux 5 et 6 qui donnent la taille de $R_D^{(\theta)}$, le critère Evaluation dont les communautés sont coûteuses à calculer, est déjà « éliminé » ; les deux critères Géographie et Age sont toujours dominants mais il y a un changement de priorité entre Motivation et Profession par rapport à la première mesure. La différence est assez négligeable pour que l'on puisse l'ignorer si l'on souhaite favoriser Profession en raison de la simplicité du calcul de $\Omega_{Profession}$ comparé au calcul de $\Omega_{Motivation}$.

Sur ces données, la mesure basée sur la consistance approximative (formule (9)) dont les résultats sont présentés dans le tableau 7, permet de départager les critères qui ne l'ont pas été par les autres mesures, et cela sans conduire à un conflit car l'ordre est compatible avec celui des autres mesures.

| | Nombre de communautés pour le critère Evaluation ($\|\Omega_{Evaluation}\|$) | | | | | |
|---|---|---|---|---|---|---|
| | 10 | 21 | 31 | 51 | 79 | 93 |
| *Géographie* | - | 1 | 3 | 4 | 4 | 4 |
| *Age* | - | 1 | 1 | 3 | 4 | 4 |
| *Motivation* | - | - | - | 2 | 3 | 3 |
| *Profession* | - | - | - | 1 | 3 | 3 |
| *Contenu* | - | - | - | - | 2 | 2 |
| *Evaluation* | - | - | - | - | - | - |

**TAB. 5** – *Analyse du nombre de réductions approximatives $R_D(\theta)$ ($\theta = 0,7$ et $\alpha = 1$)*

| | Nombre de communautés pour le critère Evaluation ($\|\Omega_{Evaluation}\|$) | | | | | |
|---|---|---|---|---|---|---|
| | 10 | 21 | 31 | 51 | 79 | 93 |
| *Age* | - | 1 | 1 | 1 | 2 | 3 |
| *Motivation* | - | - | - | 1 | 1 | 1 |
| *Profession* | - | - | - | 1 | 1 | 1 |
| *Contenu* | - | - | - | - | 1 | 1 |
| *Evaluation* | - | - | - | - | - | - |

**TAB. 6** – *Analyse du nombre de réductions approximatives $R_D(\theta)$ ($\theta = 0,7$ et $C_0$={Géographie}).*

| | Nombre de communautés pour le critère Evaluation ($\|\Omega_{Evaluation}\|$) | | | | | |
|---|---|---|---|---|---|---|
| | 10 | 21 | 31 | 51 | 79 | 93 |
| *Géographie* | - | 72,75 | 75,19 | 80,62 | 81,04 | 82,38 |
| *Age* | - | 71,79 | 78,05 | 78,15 | 80,51 | 81,07 |
| *Profession* | - | - | - | 74,23 | 79,53 | 81,51 |
| *Motivation* | - | - | - | 73,49 | 79,39 | 81,48 |
| *Contenu* | - | - | - | - | 72,48 | 75,66 |
| *Evaluation* | - | - | - | - | - | - |

**TAB. 7** – *Analyse de la consistance approximative (%) avec $\theta = 0,7$*

Pour conclure, outre les éclairages qu'elles procurent sur les données elles-mêmes, nous constatons que ces mesures de qualité de l'attribut de décision permettent, sur ce jeu de données réelles, de proposer un ordre dans lequel traiter les critères comme attribut de décision dans le processus de correction des vecteurs de positionnement tout en tenant compte des caractéristiques des données considérées.

## 5. CONCLUSION ET PERSPECTIVES

Dans cet article, nous avons proposé le modèle des espaces de communautés fondé sur la théorie des ensembles d'approximation, ainsi que des extensions de cette théorie conduisant à des mesures de qualité pour le choix du critère de décision. Notre modèle permet d'étendre les fonctionnalités des systèmes de filtrage hybride, en les rendant capables de gérer des communautés multicritères explicites. Il permet aussi de mieux exploiter les différents critères selon la situation rencontrée en mesurant leur qualité en tant que critère clé ou décision. En effet, en se basant sur ce modèle, on peut construire une méthode d'exploitation des données « disponibles à froid » pour améliorer l'intégration de nouveaux utilisateurs dans le système [17]. De plus, notre modèle combiné avec l'utilisation des cartes de communautés [16] permettra à terme d'enrichir l'interaction avec les utilisateurs afin de surmonter les problèmes classiques de l'exploitation des systèmes de recommandation.

Nous avons mis en œuvre et testé notre modèle au travers d'expérimentations sur le jeu données réelles MovieLens.

Dans les travaux futurs, nous envisageons de varier les mesures de qualité de l'attribut de décision tout en restant dans le cadre de la théorie des ensembles d'approximation, notamment en utilisant non seulement les règles certaines dans les régions positives mais aussi les règles possibles [16], [17], [18]. Ces règles pourraient donner aux utilisateurs la possibilité de découvrir des communautés potentiellement intéressantes dans une optique exploratoire.

Nous aimerions aussi étudier la notion de « borne inférieure paramétrée » [22], qui pourrait être utilisée dans les systèmes dont la majorité des critères ne donnent pas, en tant qu'attribut de décision, des tables de communautés de haute consistance. Le principe est qu'une règle $u$ est « quasi certaine » ssi sa classe d'équivalence par $P$, $[u]_P$, est « presque » incluse dans sa classe d'équivalence par $D$, $[u]_D$ (voir 2.2.1). Cette notion permettrait de contrôler la taille des régions positives dans le but de rendre plus discriminante la consistance des tables.

Plus généralement, nous souhaitons poursuivre nos expérimentations au-delà des données de MovieLens, en mettant en œuvre notre modèle dans un système de recommandation complet afin d'évaluer l'impact de cette approche sur la qualité et la diversité des recommandations, travaux déjà entamés et décrits dans [17].

## 6. REMERCIEMENTS

Cette recherche a été partiellement soutenue par le Ministère Délégué à la Recherche et aux Nouvelles Technologies, dans le programme ACI Masses de Données, projet #MD-33.

## 7. REFERENCES

[1] R. Agrawal, A. Gupta et S. Sarawagi. Modeling Multidimensional Databases. *Proceedings of the 13th International Conference on Data Engineering (ICDE '97)*, pages 232-243, UK, 1997.

[1] M. Bouzeghoub et D. Kostadinov. Personnalisation de l'information : Aperçu de l'état de l'art et définition d'un modèle flexible de profils. *Actes de la 2ème COnférence en Recherche d'Information et Applications (CORIA'05)*, pages 201-218, Grenoble, France, 2005.

[2] J. S. Breese, D. Heckerman et C. Kadie. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. *Proceedings of the 14th Conference on Uncertainty in AI*, pages 43-52**,** USA, 1998.

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

[4] J. D. Carroll et P. Arabie. Multidimensional Scaling. *Annual Review of Psychology*, vol 31, pages 607-649, 1980.

[5] S. Coppock et L. Mazlack. Rough Sets Used in the Measurement of Similarity of Mixed Mode Data. *Proceedings of the 22th Conference of the North American Fuzzy Information Processing Society*, USA, 2003.

[6] J. W. Grzymala-Busse et M. Hu. A Comparison of Several Approaches to Missing Attribute Values in Data Mining. *Proceedings of the 2nd Conference on RS and Current Trends in Computing*, Canada, 2000.

[7] J. Han et M. Kamber. Data mining: Concepts and Techniques. *New York: Morgan-Kaufman*, 2000.

[8] J. Handl, J. Knowles et M. Dorigo. On the performance of ant-based clustering. *Proceedings of the 3rd International Conference on Hybrid Intelligence Systems*, Australia, 2003.

[9] J. L. Herlocker, J. A. Konstan et J. Riedl. Explaining Collaborative Filtering Recommendations. *Proceedings of the ACM 2000 Conference on Computer Supported Cooperative Work (CSCW'2000)*, pages 241-250, USA, 2000.

[10] A. K. Jain, M. N. Murty et P. J. Flynn. Data Clustering: A Review. *ACM Computing Surveys*, vol 31 (3), pages 264-323, 1999.

[11] D. Maier. The Theory of Relational Databases. *Computer Science*, 1983.

[12] L. Mazlack. Softly Focusing On Data. *Proceedings of the 18th Conference of the North American Fuzzy Information Processing Society (NAFIPS'99)*, USA, 1999.

[13] B. J. Mirza, B. J. Keller et N. Ramakrishnan. Studying Recommendation Algorithms by Graph Analysis. *Journal of Intelligent IS*, *vol 20 (2)*, 2003.

[14] M. Montaner, B. López et J. L. De La Rosa, A Taxonomy of Recommender Agents on the Internet. *Artificial Intelligence Review*, vol 19, Kluwer Publishers, pages 285-330, 2003.

[15] MovieLens, http://www.grouplens.org/.

[16] A-T. Nguyen, N. Denos et C. Berrut. Cartes de communautés pour l'adaptation interactive de profils dans un système de filtrage. *Actes du 32ème Congrès INFORSID*, France, 2005.

[17] A-T. Nguyen, N. Denos et C. Berrut. Exploitation des données « disponibles à froid » pour améliorer le démarrage à froid dans les systèmes de filtrage d'information. *Actes du 33ème Congrès INFORSID*, Tunisie, 2006.

[18] S-H. Nguyen et H-S. Nguyen. Some efficient algorithms for rough set methods. *Proceddings of the Conference of Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Granada, Spain, pages 1451-1456, 1996.

[19] Z. Pawlak. Rough Sets. *International Journal of Computer and Information Sciences*, vol 11 (5), pages 341-356, 1982.

[20] Z. Pawlak. Rough Classification. *International Journal of Man-Machine Studies*, vol 20, pages 469-483, 1984.

[21] Z. Pawlak. Some Issues on Rough Sets. *Transaction on Rough Sets I, LNCS 3100*, 2004.

[22] S. Perugini, M. A. Gonçalves et E. A. Fox. A Connection-Centric Survey of Recommender Systems Research. *Journal of Intelligent IS*, vol 23 (1), 2003.

[23] L. Polkowski. Rough Sets: Mathematical Foundations. *Physica-Verlag*, 2002.

[24] A. M. Rashid, I. Albert, D. Cosley, S. K. Lam, S. M. Mcnee, J. A. Konstan et J. Riedl. Getting to Know You: Learning New User Preferences in Recommender Systems. *Proceedings of the 7th International Conference on Intelligent User Interfaces (IUI'02)*, California, USA, pages 127-134, 2002.

[25] A. Skowron et C. Rauszer. The Discernibility Matrices and Functions in Information Systems. *Intelligent Decision Support: Handbook of Applications and Advances of RS Theory, Series: Theory and Decision Library*, vol 11, Kluwer Academic Publishers, pages 331-362, 1992.

[26] M. Zhang et J. T. Yao. A Rough Sets Based Approach to Feature Selection. *Proceedings of the 23rd Conference of the North American Fuzzy Information Processing Society*, Canada, 2004.