Modèle d'espaces de communautés basé sur la théorie des ensembles d'approximation dans un système de filtrage hybride

An-Te Nguyen, Nathalie Denos, Catherine Berrut

Modèle d'espaces de communautés basé sur la théorie des ensembles d'approximation dans un système de filtrage hybride

An-Te Nguyen, Nathalie Denos, Catherine Berrut (Chercheurs)

Laboratoire CLIPS-IMAG
385, rue de la Bibliothèque
BP 53, 38041 Grenoble Cedex
{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 un modèle d'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. Le modèle d'espaces de communautés permet de diversifier les recommandations qui peuvent émaner de communautés variées. Les mesures permettent de comparer des critères entre eux afin de déterminer une priorité entre les critères dans la tâche d'amélioration du positionnement des utilisateurs dans les communautés.

ABSTRACT. *Recommender systems intend to provide relevant information based on user preference. Most of them rely on collaborative filtering that compare users on the basis of their ratings in order to group them into communities and then produce recommendations on this basis. We propose a model that allows to explicitly manage multi criteria community spaces in recommender systems, in order to diversify recommendations from different communities. We also define several 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.*

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

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

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.

1. Introduction

L'objectif des systèmes de filtrage adaptatif, ou systèmes de recommandation 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 (Montaner *et al.*, 03). Les moteurs de ces systèmes gèrent des profils d'utilisateurs permettant de sélectionner les recommandations, et adaptent ces profils en exploitant le retour de pertinence fourni par les utilisateurs.

Il existe deux grandes approches de filtrage : basé sur le contenu et collaboratif. Le filtrage basé sur le contenu compare les nouveaux documents au profil de chaque utilisateur et recommande ceux qui sont le plus proche (Montaner *et al.*, 03). 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 (Breese *et al.*, 98). Le filtrage hybride combine ces deux approches (Burke, 02) : en général les profils sont orientés contenu, et la comparaison entre ces profils donne lieu à la formation de communautés permettant le filtrage collaboratif. Dans ce domaine, la notion de *communauté* joue un rôle clé. Nous abordons dans la suite la question de leur formation et de leur exploitation.

Dans la plupart des systèmes de filtrage[1], les communautés sont formées selon un seul critère (Perugini *et al.*, 03), comme la proximité des évaluations des utilisateurs. Pourtant on peut appuyer la formation de communautés sur de multiples critères. Par exemple, dans un système de recommandation de films le profil de l'utilisateur u_3 (voir **Figure 1**) contient sa profession (Chercheur), la ville où il habite (Paris), son genre de film préféré (Documentaire) et ses évaluations sur certains films (ce en quoi il appartient au Groupe 2 d'utilisateurs, dont les évaluations sont proches). Pour chacun de ces attributs, on peut regrouper les utilisateurs par leur proximité relativement à cet attribut ou critère. Ainsi, u_3 est proche de u_4 pour le critère Profession mais éloigné pour le critère Genre préféré. Ainsi, en plus de l'espace de communautés créé par le critère Evaluation, le système peut construire un espace pour chacun des critères Profession, Ville et Genre préféré. La **Figure 1** montre les espaces de communautés ainsi construits.

Dans ce contexte, la question se pose de quel critère utiliser pour former les communautés servant de base aux recommandations. Nous avons montré (Nguyen *et al.*, 05) que les espaces de communautés formés sur les critères Genre préféré et Evaluation diffèrent, d'où l'idée d'enrichir la production de recommandations, un utilisateur pouvant recevoir les recommandations par filtrage collaboratif en provenance de chacune des communautés auxquelles il appartient. Ainsi l'utilisateur u_3 peut aussi recevoir les films envoyés via les communautés Chercheur, Paris et Documentaire outre ceux du critère Evaluation (filtrage collaboratif classique). Cette diversité des recommandations peut également être exploitée de façon sélective dans une stratégie adaptée à la situation des utilisateurs.

[1] Dans le reste de l'article, le terme « système de filtrage » désigne un système hybride.

Critère / 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

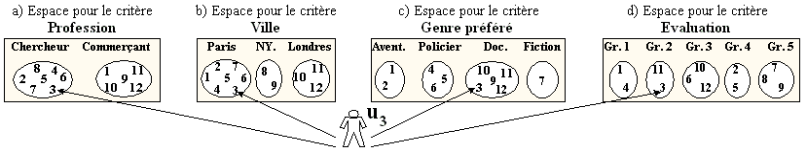

Figure 1. *Table de communautés d'un système de recommandation de films et espaces de communautés associés à chaque critère.*

La qualité du positionnement des utilisateurs dans les espaces dépend fondamentalement de la qualité des valeurs données pour chaque utilisateur à chaque critère. Certains critères demandent beaucoup d'efforts aux utilisateurs, et peuvent être coûteux également pour le système (Montaner *et al.*, 03). 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 fastidieuse, et c'est pourtant sur cette base que les communautés sont formées pour le critère Evaluation. Cela conduit à l'absence de valeur pour certains critères (p. ex. on ne connaît pas le genre préféré) et à des valeurs douteuses (p. ex. les évaluations sont en très petit nombre), d'où un positionnement impossible ou peu fiable des utilisateurs. Nous contribuons ici à résoudre ce problème, et posons les bases pour élaborer une stratégie d'amélioration des positionnements, et donc des profils.

Nous formalisons d'abord 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 sur tout critère de rapprochement entre utilisateurs. Un utilisateur appartient à une communauté dans chaque espace, ce qui est traduit par son « vecteur de positionnement ». Nous analysons ensuite les relations entre les critères, grâce à des mesures évaluant leurs capacités à se déterminer les uns les autres (p. ex. Ville = New York et Genre = Documentaire déterminent Evaluation = Groupe 5), compte tenu desquels sont prioritaires dans le système (p. ex. on connaît toujours Profession et Ville mais pas Genre et Evaluation). Nous *ordonnons alors ces critères* pour servir de base aux stratégies d'*amélioration des vecteurs de positionnement*, et à terme améliorer les recommandations.

Nous formalisons les espaces de communautés avec la *théorie des ensembles d'approximation* (« rough sets ») (Pawlak, 04). Cette théorie permet aussi d'analyser la dépendance entre un critère clé (ou décision) fixé *a priori* et les autres critères. Nous en proposons une extension avec des mesures de comparaison des critères sans fixer de décision *a priori*. Puis nous validons le modèle sur le jeu de données MovieLens.

2. Contexte et état de l'art

2.1. *Notion de communauté*

Dans (Perugini *et al.*, 03), les communautés sont étudiées sous l'angle « social » de l'effort à produire pour établir des relations entre utilisateurs. Ces communautés sont étudiées dans (Montaner *et al.*, 03) sous l'angle fonctionnel. Il ressort des travaux existants que les communautés sont formées sur un unique critère, demeurant implicite, 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. (Bouzeghoub *et al.*, 05). En utilisant chacun de ces critères, le système peut former autant d'*espaces de communautés* (voir **Figure 1**) où positionner un même utilisateur.

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

Notre problématique requiert le traitement de données nominales (valeurs des attributs), ainsi que la prise en compte et l'analyse des imperfections de ces données (valeurs manquantes, douteuses, périmées). Après une étude des théories applicables aux problèmes de données manquantes (Coppock *et al.*, 03 ; Mazlack, 99), nous avons choisi la théorie des ensembles d'approximation (Pawlak, 04 ; Polkowski, 02) car elle permet de traiter des données nominales (voir **Figure 1**) et ne requiert aucune information a priori. Elle permet aussi d'identifier et exploiter les dépendances entre les attributs des données. Ici, les objets étudiés sont les utilisateurs, leurs attributs sont les 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 compensons l'absence de certaines valeurs d'attributs en exploitant les attributs où les valeurs sont disponibles (p. ex. connaissant Ville, Profession et Genre, proposer une valeur pour Evaluation sans demander d'évaluations à l'utilisateur).

2.2.1. Représentation des données et ensembles d'approximation

Nous présentons ici les notions qui sont utiles à la compréhension du modèle. 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 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 est considérée comme une *règle de décision*, p. ex. u_7 : (Profession = Chercheur, Ville = Paris, Genre = Fiction) → (Evaluation = Groupe 5).

		P={Ville, Genre}					
Crit. / Ut.	Profession	Ville	Genre préféré	Evaluation	POS_P (Eval)	POS_C (Genre)	POS_C (Eval)
u_1	Commerçant	Paris	Aventure	Groupe 1		x (inc)	x
u_2	Chercheur	Paris	Aventure	Groupe 4			x
u_3	Chercheur	Paris	Documentaire	Groupe 2	x (inc)	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	x
u_8	Chercheur	New York	Documentaire	Groupe 5	x	x	x
u_9	Commerçant	New York	Documentaire	Groupe 5	x	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	

Tableau 1. *Ensembles d'approximation (D = {Evaluation} et P = {Ville, Genre}) puis comparaison des critères Genre et Evaluation comme décision*

Les *règles certaines* sont celles pour lesquelles un objet peut être certainement classé dans une communauté en se basant sur P. Par exemple, dans le **Tableau 1** il y a 4 règles certaines relativement à P = {Ville, Genre} pour la décision D = {Evaluation} : notamment u_3, car pour (Ville = Paris, Genre = Documentaire) on a toujours (Evaluation = Groupe 2), et de même pour u_7, u_8 et u_9.

La *région positive* de la table T pour les conditions P et la décision D, notée $POS_P(D)$, est l'ensemble de toutes les règles certaines de T : $POS_P(Eval) = \{u_3, u_7, u_8, u_9\}$. Ce sont les objets ayant toujours la même valeur pour Evaluation étant donnée une valeur pour Ville et pour Genre.

2.2.2. Dépendance d'une décision, consistance de la table et signification de conditions

La « dépendance » de la décision D par rapport à l'ensemble $P \subseteq C$ traduit la part des objets donnant lieu à des règles certaines, et peut être mesurée par :

$$\gamma(P, D) = \frac{|POS_P(D)|}{|U|} \qquad [1]$$

Pour l'ensemble des conditions C, $\gamma(C,D)$ exprime la « consistance » de la table T vis-à-vis de la décision D par les conditions C.

La réduction de C est le plus petit sous-ensemble $P \subseteq C$ préservant la région positive : $POS_P(D) = POS_C(D)$. Une même table peut donner plusieurs réductions.

L'importance d'un sous-ensemble de conditions $P \subset C$ est traduite par sa « signification », mesurée comme l'impact de leur suppression sur la consistance :

$$\sigma_{C,D}(P) = 1 - \frac{\gamma(C \setminus P, D)}{\gamma(C, D)} \qquad \text{(appartient à [0,1])} \qquad [2]$$

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. Modèle d'espaces de communautés

Nous présentons ici notre modèle basé sur la théorie des ensembles d'approximation. Les informations de la table de communautés T sont obtenues à partir des données disponibles dans les profils des utilisateurs. Certains critères sont obtenus de façon directe (ville) ; pour certains autres on doit regrouper plusieurs valeurs (tranches d'âge) pour former des communautés plus significatives ; pour d'autres on doit faire appel à une méthode de classification (Jain *et al.*, 99) en utilisant une mesure de similarité (évaluation). Nous avons montré dans (Nguyen *et al.*, 05) 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. Enfin nous définissons les mesures destinées à comparer entre eux les critères.

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 Ω_a, est appelée « espace de communautés pour a ». Elle 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 ».

En généralisant cette formalisation, on définit la notion de « critère composé » P comme un ensemble d'au moins deux critères $P \subset A$, et l'« espace de communautés composé » Ω_P est 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 » $\mathcal{P}_u$ de l'utilisateur u la liste des étiquettes de ses propres communautés selon chacun des critères (il correspond à la ligne T[u]). Ces vecteurs traduisent le polymorphisme du positionnement des utilisateurs au sein des communautés : 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 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 le cas d'un vecteur où il ne manque qu'une valeur, on prend comme décision D le critère correspondant à la valeur manquante, et on cherche une règle certaine dans la région positive avec P = A \ D afin d'inférer la valeur pour D. Supposons que le nouvel utilisateur a pour vecteur de positionnement (Commerçant, Paris, Aventure, _), sa communauté pour Evaluation n'étant pas encore connue car il n'a évalué aucun film. Grâce à la règle (Profession = Commerçant, Ville = Paris, Genre = Aventure) → (Evaluation = Groupe 1), le système peut rattacher cet utilisateur à la communauté libellée Groupe 1. Si D dépendait totalement des critères restants, on trouverait toujours une telle règle pour compléter la valeur pour D. Mais ce n'est pas souvent le cas, et on doit faire appel à des techniques existantes de remplissage de valeurs manquantes, comme par exemple l'approche probabiliste, techniques que nous ne détaillons pas ici (Grzymala-Busse *et al.*, 00).

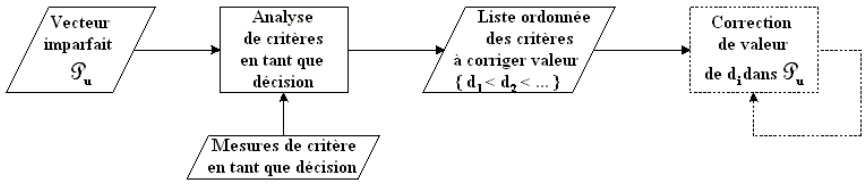

Figure 2. *Mécanisme générique de correction d'un vecteur de positionnement.*

Si plusieurs valeurs sont manquantes ou douteuses, il faut élaborer une stratégie pour décider dans quel ordre compléter ou modifier ces 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 décision » permettant de comparer les critères afin de déterminer cet ordre (voir **Figure 2**), partant du principe qu'une table de bonne consistance donnera lieu à de bonnes corrections.

3.2. *Mesures de qualité de décision*

Nous proposons trois mesures qui s'appuient sur la consistance de la table et la signification de conditions.

Mesure basée sur la consistance. Cette mesure favorise la décision qui fournit les règles certaines les plus nombreuses. Elle est basée sur la consistance $\gamma(C, D)$ de la table T pour les conditions C par rapport à la décision. On dit que la décision D_2 est meilleure que la décision D_1 si elle donne lieu à une meilleure consistance :

$$D_1 \lhd D_2 \Leftrightarrow | POS_{A\setminus D1}(D_1) | \leq | POS_{A\setminus D2}(D_2) | \quad [3]$$

Prenons l'exemple d'un utilisateur dont les valeurs des critères Evaluation et Genre manquent dans son vecteur de positionnement. Alors, le **Tableau 1** montre que Genre est une meilleur décision du fait de la plus grande taille de sa région positive, les

colonnes POS_C(Genre) et POS_C (Eval) indiquant les règles certaines incluses dans les régions positives pour les décisions Genre et Evaluation respectivement.

Mesures basées sur les réductions approximatives. Ces mesures de qualité de décision sont définies à partir des réductions approximatives afin de tenir compte de la complexité du calcul : on cherche à favoriser les décisions conduisant à un sous-ensemble P de conditions aussi réduit que possible tout en conservant une signification supérieure à un seuil θ. Etant donné la décision D et les conditions C, on définit d'abord les réductions approximatives $\mathbf{R}_D^{(\theta)}$, étant donné le seuil θ :

$$\mathbf{R}_D^{(\theta)} = \{P \subseteq C \mid \sigma(P) \geq \theta\} \quad [4]$$

Ce sont les sous-ensembles de critères « acceptables » et inclus dans C. Par exemple, si θ est égal à 0,8, $\mathbf{R}_{Eval}^{(\theta)}$ contient deux réductions approximatives : P_1 = {Ville, Genre} et P_2 = {Profession, Ville}.

A partir des ensembles $\mathbf{R}_D^{(\theta)}$ on peut définir diverses relations d'ordre sur les décisions, soit en fixant un nombre maximum α de conditions à conserver, soit en fixant un ensemble C_0 des conditions jugées utiles dans un contexte donné :

$$D_1 \lhd D_2 \Leftrightarrow |\{P \in \mathbf{R}_{D1}^{(\theta)} \mid |P| \leq \alpha\}| \leq |\{Q \in \mathbf{R}_{D2}^{(\theta)} \mid |Q| \leq \alpha\}| \quad [5]$$

$$D_1 \lhd D_2 \Leftrightarrow |\{P \in \mathbf{R}_{D1}^{(\theta)} \mid C_0 \subseteq P\}| \leq |\{Q \in \mathbf{R}_{D2}^{(\theta)} \mid C_0 \subseteq Q\}| \quad [6]$$

La formule [5] favorise la décision qui donne de petites réductions, ce qui signifie que l'on a besoin de connaître peu de choses sur l'utilisateur. Plutôt que la taille des réductions, la formule [6] prend en compte leur contenu, préférant les conditions demandant peu d'effort à produire par l'utilisateur, p. ex. son code postal.

En pratique, puisque la tâche de déterminer toutes les réductions relatives à une décision D (nécessaire pour analyser sa qualité par [5] ou [6]) est un problème NP-complet, on peut limiter a priori la taille de $\mathbf{R}_D^{(\theta)}$ pour diminuer la complexité du calcul, ou calibrer le seuil θ pour que les réductions dans $\mathbf{R}_D^{(\theta)}$ contiennent dans la plupart des cas les critères de C_0.

Mesure basée sur la consistance approximative. Elle permet de départager les décisions qui ne le sont pas par les mesures [5] ou [6]. Elle repose sur la consistance μ de la table de communautés par rapport aux réductions approximatives.

$$\mu(D) = \frac{1}{\left|R_D^{(\theta)}\right|} \sum_{P \in R_D^{(\theta)}} \sigma(P) \quad [7]$$

$$D_1 \lhd D_2 \Leftrightarrow \mu(D_1) \leq \mu(D_2) \quad [8]$$

3.3. *Utilisation des mesures pour corriger un vecteur de positionnement*

Supposons que pour un nouvel utilisateur u les valeurs des critères Genre et Evaluation dans le vecteur de positionnement ne sont pas encore connues :

$\mathcal{P}_u$ = (Commerçant, Londres, _, _). D'abord, le système choisit le critère Genre comme décision en utilisant la mesure basée sur la consistance (voir **Figure 2**). Ensuite, les « règles certaines » u_{10}, u_{11} et u_{12} permettent d'inférer Documentaire comme la valeur du critère Genre. Alors, on obtient : $\mathcal{P}_u$ = (Commerçant, Londres, Documentaire, _). Enfin, le critère Evaluation pourrait être instancié par Groupe 3 qui est la valeur dominante dans les trois règles concernées (Grzymala-Busse *et al.*, 00). 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 d'utilisateurs.

Nous soulignons ici une différence importante entre notre modèle et d'autres applications de la théorie des ensembles d'approximation : la théorie fait l'hypothèse que la décision est fixée dès la conception du système, et que l'on essaie de sélectionner à partir d'un ensemble d'apprentissage des réductions de conditions en conservant la qualité de la table de décision. On peut dire que cette théorie focalise sur la qualité des conditions plutôt que des décisions. Au contraire, dans notre modèle tous les critères sont sur le même plan, et pendant l'exploitation le système va choisir parmi les critères la meilleure 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, notre modèle ne vise pas à remplacer les systèmes de filtrage existants, mais offre un cadre plus large pour les mettre en œuvre (communautés multicritères) et la possibilité de repositionner les utilisateurs en situation problématique.

4. Validation du modèle dans un système de recommandation de films

Ces travaux de validation montrent comment utiliser ce modèle d'espaces de communautés dans un système de filtrage. Nous construisons d'abord la table de communautés pour un système réel (MovieLens) qui dispose de données variées. Nous analysons ensuite les critères en utilisant les mesures proposées plus haut[2].

4.1. *Construction de la table de communautés*

Le jeu de données MovieLens 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. La table de communautés sera composée des 943 vecteurs de positionnement avec 6 colonnes : Age, Profession, Géographie, Motivation, Contenu et Evaluation. Pour les 4 premières, la création des espaces Ω est simple et le nombre de communautés fixe dans chaque Ω :

[2] Bien que le calcul des ensembles d'approximation soit en général un problème NP-complet, nous n'avons utilisé aucune méthode heuristique (Zhang *et al.*, 04) en raison du nombre faible de critères et du temps de calcul raisonnable (une dizaine de minutes par analyse).

– 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 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 (Nguyen *et al.*, 05) : la méthode des fourmis artificielles (Handl *et al.*, 03) place les utilisateurs dans un espace en 2D, puis une classification ascendante hiérarchique (Jain *et al.*, 99) permet d'obtenir un nombre de communautés flexible.

En pratique, les utilisateurs sont souvent bien regroupés dans $\Omega_{Contenu}$, et assez dispersés dans $\Omega_{Evaluation}$ (Nguyen *et al.*, 05) en raison du faible nombre d'objets jugés en commun (Breese *et al.*, 98). 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, ce pourquoi les expériences ont été paramétrées par le nombre de communautés dans $\Omega_{Evaluation}$.

4.2. *Résultats et analyse*

Nous cherchons à établir un ordre entre les critères, pour savoir lequel utiliser comme critère clé pour 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 obtiennent la priorité la plus faible comme critère clé, alors qu'ils correspondent aux communautés les plus coûteuses à former. On peut donc envisager de limiter ce coût en positionnant les nouveaux utilisateurs dans ces deux espaces via leurs positions dans les autres espaces. Cette approche améliore la performance de la formation multiple de communautés.

Le **Tableau 2a** consacré à la mesure [3] montre que le critère Evaluation est la pire décision, que Géographie et Age sont les meilleurs, et que Profession et Motivation viennent ensuite, avant le critère Contenu. Plus le nombre de communautés de $\Omega_{Evaluation}$ diminue (de 93 à 10), plus la consistance de la table se dégrade excepté celle du critère Evaluation.

Pour la mesure [5], le paramètre α peut varier de 1 à 5 selon le nombre de critères souhaité dans les conditions. 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 θ a été choisie en se basant sur le tableau

a. Pour [6], Géographie pour C_0 est simplement un exemple. Dans les tableaux **b** et **c** qui donnent $|\mathbf{R}_D^{(\theta)}|$, le critère Evaluation est déjà « éliminé » ; les critères Géographie et Age dominent encore mais Motivation et Profession changent de priorité par rapport à la première mesure. La différence est assez négligeable pour qu'on puisse l'ignorer afin de favoriser Profession pour la simplicité de calcul des communautés.

$\lvert\Omega_{Eval}\rvert$ / Critères	**a** : consistance [3]					
	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

$\lvert\Omega_{Eval.}\rvert$ / Critères	**b** : réductions app. [5]					
	10	21	31	51	79	93
Géographie	-	1	3	4	4	4
Age	-	1	1	3	4	4
Profession	-	-	-	2	3	3
Motivation	-	-	-	1	3	3
Contenu	-	-	-	-	2	2

$\lvert\Omega_{Eval}\rvert$ / Critères	**c** : réductions app. [6]					
	10	21	31	51	79	93
Age	-	1	1	1	2	3
Motivation	-	-	-	1	1	1
Profession	-	-	-	1	1	1
Contenu	-	-	-	-	1	1
Evaluation	-	-	-	-	-	-

$\lvert\Omega_{Eval}\rvert$ / Critères	**d** : consistance app. (%) [8]					
	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	-	-	-	-	-	-

Tableau 2. *Analyse de la qualité des critères en tant que décision*

Enfin, la mesure [8] (tableau **d)**, permet de départager les critères non distingués par les autres mesures, et cela sans conduire à un conflit. Pour conclure, outre les éclairages qu'elles procurent sur les données elles-mêmes, ces mesures de qualité de décision fournissent un ordre dans lequel traiter les critères comme décision dans le processus de correction des vecteurs de positionnement, tout en tenant compte des caractéristiques des données considérées.

5. Conclusion et perspectives

Nous avons proposé un modèle d'espaces de communautés fondé sur la théorie des ensembles d'approximation, et des extensions conduisant à des mesures de qualité pour le critère de décision. Ce modèle permet d'étendre les fonctionnalités des systèmes de filtrage hybride, afin de gérer des communautés multicritères explicites, et de mieux exploiter les différents critères selon la situation rencontrée en mesurant leur qualité en tant que critère clé ou décision. Ce modèle combiné avec l'utilisation des cartes de communautés (Nguyen *et al.*, 05) permettra à terme d'enrichir l'interaction avec les utilisateurs et de contribuer à surmonter les problèmes classiques de ces systèmes.

Nous envisageons de varier les mesures de qualité de décision : p. ex. en exploitant aussi les « bornes supérieures » (Pawlak, 04), qui contiennent les règles possibles et

pourraient amener les utilisateurs à découvrir des communautés potentiellement intéressantes (optique exploratoire). Nous aimerions aussi étudier la notion de « borne inférieure paramétrée » (Zhang *et al.*, 04), pour le cas où les critères ne donnent pas des tables de communautés de haute consistance, dans le but de contrôler la taille des régions positives en vue de rendre plus discriminante la consistance des tables.

Plus généralement, nous souhaitons poursuivre nos expérimentations, 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.

6. Bibliographie

Bouzghoub M., Kostadinov D., « Personnalisation de l'information : Aperçu de l'état de l'art et définition d'un modèle flexible de profils », *CORIA'05*, France, 2005.

Breese J. S., Heckerman D., Kadie C., « Empirical Analysis of Predictive Algorithms for Collaborative Filtering », *The 14th Conference on Uncertainty In AI*, 1998**,** USA, p.43-52.

Burke R., « Hybrid Recommender Systems: Survey and Experiments », *Journal of Personalization Research, User Modeling and User-Adapted Interaction*, vol. 12 (4), 2002.

Coppock S., Mazlack L., « Rough Sets Used in the Measurement of Similarity of Mixed Mode Data », *22nd Conference of the North American Fuzzy Information Processing Society*, 2003.

Grzymala-Busse J. W., Hu M., « A Comparison of Several Approaches to Missing Attribute Values in Data Mining », *2nd Conference on RS and Current Trends in Computing*, 2000.

Handl J., Knowles J., Dorigo M., « On the performance of ant-based clustering », *The 3rd International Conference on Hybrid Intelligence Systems*, Australia, 2003.

Jain A. K., Murty M. N., Flynn P. J., « Data Clustering: A Review », *ACM Computing Surveys*, vol. 31 (3), 1999, p.264-323.

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

Montaner M., López B., De La Rosa J. L., « A Taxonomy of Recommender Agents on the Internet », *Artificial Intelligence Review*, vol. 19, 2003, Kluwer Publishers, p.285-330.

MovieLens, http://www.grouplens.org/.

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

Pawlak Z., « Some Issues on Rough Sets », *Transaction on Rough Sets I, LNCS 3100*, 2004.

Perugini S., Gonçalves M. A., Fox E. A., « A Connection-Centric Survey of Recommender Systems Research », *Journal of Intelligent Information Systems*, vol. 23 (1), 2003.

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

Zhang M., Yao J. T., « A Rough Sets Based Approach to Feature Selection », *The 23rd Conference of the North American Fuzzy Information Processing Society*, Canada, 2004.