Un modèle de recommandation contextuel pour la prédiction des intérêts des consommateurs sur le Web

Mohamed Ramzi Haddad[1], Hajer Baazaoui[1], Djemel Ziou[2], Henda Ben Ghezala[3]

[1] LABORATOIRE RIADI-GDL, École Nationale des Sciences de l'Informatique, Université de la Manouba, La Manouba 2010, Tunisie
haddad.medramzi@gmail.com, hajer.baazaouizghal@riadi.rnu.tn, henda.benghezala@riadi.rnu.tn

[2] Centre de recherche MoIVRe, Département d'informatique, Faculté des sciences, Université de Sherbrooke, 2500 Boul. Université, Sherbrooke, J1K 2R1, Canada
djemel.ziou@usherbrooke.ca

Résumé : Avec l'explosion du commerce sur le Web, il devient difficile de cibler les besoins des consommateurs avec les produits qui conviennent le plus à leurs préférences. Dans cet article, nous proposons et évaluons un modèle de recommandation probabiliste ayant pour objectif de prédire les intérêts et les achats des consommateurs. Le modèle de recommandation proposé tient compte des connaissances sur la psychologie du consommateur et intègre les différents facteurs qui influencent les comportements de consommation tels que la démographie, les caractéristiques des produits, les évaluations, le contexte et l'historique des achats. L'expérimentation comparative montre que notre modèle unifie les principales idées directrices des approches classiques et donne de meilleurs résultats sur un jeu de données réelles.

Mots-clés : recommandation comportementale, filtrage contextuel, modélisation du consommateur, prédiction des intérêts et des achats

1 Introduction

Les consommateurs sont souvent confrontés à un grand nombre d'alternatives et d'informations prêtant les moins expérimentés à la confusion. Dans les galeries marchandes classiques, les vendeurs avaient comme responsabilité de déterminer les besoins et les préférences des clients afin de mieux les conseiller et garantir leur satisfaction par le produit à acquérir. Dans le cas d'un achat en ligne, les consommateurs ne sont plus guidés ni conseillés en explorant les offres mises à leurs dispositions. Le consommateur peut faire face au problème de la paralysie par l'analyse (Iyengar & Lepper, 2000) et ne pas prendre une décision face à une surcharge d'information et de possibilités, surtout lorsque celui-ci n'a pas de connaissance sur le produit ou lorsque le coût d'acquisition est élevé. Par ailleurs, selon (Schwartz, 2005), plus les consommateurs ont de choix concernant un type donné de biens, plus ils seraient susceptibles de regretter les décisions et les choix qu'ils vont prendre.

Plusieurs études sur la psychologie du consommateur ont confirmé l'existence du problème de surcharge de choix et ses implications sur le bien-être des consommateurs (Schwartz, 2005; Jacoby *et al.*, 1974; Iyengar & Lepper, 2000). De plus, ils insistent sur la nécessité de réduire les choix mis à disposition des consommateurs afin de permettre et d'accélérer leur prise de décision (Schwartz, 2005). Cette solution reviendrait alors à filtrer les ressources proposées

au consommateur. Plusieurs propositions ayant pour objectif le filtrage de l'information ont été proposées comme les approches de personnalisation et de recommandation. Ces approches cherchent à cibler les consommateurs avec les produits ou le contenu les plus appropriés et limiter ainsi les effets de la complexité de l'information et de la surcharge de choix.

Le modèle de recommandation que nous proposons dans cet article s'inspire des travaux sur la psychologie du consommateur et a pour objectif de prédire les intérêts et les achats de celui-ci. Notre méthodologie consiste à prendre en considération tous les facteurs qui influenceraient les intérêts et les décisions d'achat du consommateur dans l'optique de généraliser les idées et les hypothèses directrices des approches de recommandation existantes. Pour cela, le modèle se base sur plusieurs facteurs, à savoir (1) la démographie, (2) les attitudes des consommateurs, (3) les propriétés des ressources (p. ex. produits, services, contenu informationnel, etc...), (4) le contexte de consommation et (5) les historiques des achats. Il s'agit donc de recenser et de formaliser ces facteurs afin d'étudier leurs causalités et mieux prédire les comportements de consommation moyennant un modèle statistique.

Cet article est organisé comme suit. Dans la section 2, nous présentons un état de l'art des approches recommandation. Ensuite, dans la section 3, nous définissons les facteurs retenus qui influenceraient les décisions des consommateurs afin de parvenir à la modélisation de l'approche recommandation. Par la suite, nous formalisons le modèle de recommandation proposé et détaillons la méthodologie de génération des recommandations. Enfin, dans la section 4, nous présentons une expérimentation comparative entre plusieurs variantes de notre modèle et des approches de recommandation classiques afin de mesurer et de discuter l'apport de notre proposition sur un jeu de données réelles.

2 État de l'art des approches de recommandation

Les approches de recommandation adoptent des idées directrices et des hypothèses différentes afin de déterminer les ressources les plus pertinentes pour un utilisateur donné. Par conséquence, selon leurs processus d'inférence et de génération des recommandations, les approches existantes peuvent être classées parmi les catégories suivantes.

2.1 Approches basées sur les attributs

Ces approches évaluent la pertinence des ressources sur la base de leurs attributs. En effet, plus les attributs d'une ressource concordent avec les préférences des utilisateurs, plus elle est susceptible d'être recommandée. Les principales approches basées sur les attributs sont les approches de filtrage par contenu (CBF) (Balabanović & Shoham, 1997; Mooney & Roy, 2000).

2.2 Approches basées sur les corrélations des utilisateurs

La génération des recommandations est basée sur les corrélations qui peuvent exister entre les utilisateurs. Dans ce cadre, les utilisateurs corrélés peuvent être ceux ayant des ressources préférées communes, des patrons de notation concordants (Resnick *et al.*, 1994; Goldberg *et al.*, 1992) ou des attributs démographiques similaires (Krulwich, 1997; Pazzani, 1999). Les recommandations proposées à un utilisateur regroupent les ressources ayant été jugées comme pertinentes par les utilisateurs qui lui sont corrélés. D'autres variantes tentent de prédire les notes

que donnerait l'utilisateur aux ressources qu'il ne connaît pas afin de ne retenir que celles qui seraient les plus intéressantes. Le filtrage collaboratif orienté utilisateur (UC-CF) et le filtrage démographique (DF) utilisent ces types de raisonnements afin de générer des recommandations.

2.3 Approches basées sur les corrélations des ressources

Les recommandations sont générées à partir des corrélations entre les ressources disponibles et celles déjà préférées ou adoptées par l'utilisateur. Par exemple, deux produits sont considérés comme corrélés s'ils sont fréquemment achetés ensemble ou jugés comme intéressants par les mêmes groupes d'utilisateurs (Linden *et al.*, 2003). Les recommandations issues sont souvent expliquées par le constat suivant : "les utilisateurs qui ont apprécié/acheté ce produit on aussi apprécié/acheté les produits A et B". D'autres approches utilisent les corrélations entre les notes obtenues par les ressources afin de déterminer les produits ayant les mêmes patrons de notations (Sarwar *et al.*, 2001; Deshpande & Karypis, 2004). Ensuite, la note qu'affecterait un utilisateur à un produit est estimée en se basant sur les notes qu'il aurait données aux ressources qui lui sont corrélées. Le filtrage collaboratif orienté item (IC-CF) et les algorithmes à base de règles d'association utilisent ces types de raisonnement pour générer les recommandations.

2.4 Approches basées sur les connaissances

Les recommandations sont issues d'un processus d'inférence encapsulant des connaissances du domaine et permettant de déduire les ressources pertinentes à partir des besoins et des préférences des utilisateurs (Burke, 2000; Lin *et al.*, 2002). Ces connaissances associent les besoins et les contraintes de l'utilisateur aux ressources pouvant les satisfaire. Ces connaissances peuvent être déterminées par des experts, exprimées explicitement par les consommateurs ou apprises à partir des données disponibles grâce aux techniques de fouille de données (Agrawal & Srikant, 1994). Cette classe d'approches se distingue par la représentation formelle et fonctionnelle des connaissances permettant le raisonnement et la déduction des recommandations grâce à des mécanismes d'inférence (p.ex. représentations à base de règles et à base de cas).

2.5 Approches contextuelles

Des recherches sur la personnalisation et la recommandation ont souligné l'influence du contexte sur les décisions et les intérêts des consommateurs. Ceci a conduit à des approches de recommandation sensible au contexte (Woerndl & Groh, 2007; Jones, 2005). De telles approches sont principalement dérivées des approches existantes augmentées par la dimension contextuelle afin d'identifier les cas où le contexte, tel que le cadre spatio-temporel ou le support d'accès à l'information, implique certains comportements de consommation communs et prévisibles (Boutemedjet & Ziou, 2008; Baazaoui *et al.*, 2014).

2.6 Approches hybrides et comportementales

Plusieurs recherches récentes s'intéressent à la composition et à l'utilisation simultanée de plusieurs approches de recommandation dans des approches hybrides. Ces travaux partent de l'hypothèse que l'hybridation des approches de recommandation pourrait améliorer leur efficacité et permettrait de surmonter leurs lacunes. En effet, les systèmes de recommandation peuvent

être vu comme des classifieurs dont le rôle est de séparer les produits pertinents des autres et peuvent donc bénéficier des techniques d'agrégation ou de composition de classifieurs. Ce type de systèmes hybrides repose donc sur un ensemble de sous-systèmes "naïfs", mais spécialisés, dont l'agrégation améliore la qualité de la recommandation. Dans (Burke, 2007), l'auteur justifie le recours aux systèmes de recommandation hybrides et définit plusieurs stratégies d'hybridation selon le contexte et les besoins du domaine. Enfin, plusieurs recherches s'intéressent à l'analyse des réseaux sociaux et des sentiments des consommateurs afin de mieux cibler leurs préférences et anticiper leurs comportements (hsien Liao *et al.*, 2012; Poussevin *et al.*, 2014).

2.7 Discussion et objectifs

Bien que les études sur les comportements de consommation ont recensé plusieurs facteurs qui influencent les décisions des consommateurs, les approches de recommandation existantes sont basées sur des hypothèses et des constatations simples et réductrices sur la façon avec laquelle les consommateurs font leurs choix (Jacoby *et al.*, 1974; Iyengar & Lepper, 2000; Hawkins *et al.*, 2003; Schwartz, 2005). En effet, les approches de recommandation hybrides ont démontré que la composition de plusieurs approches permettait d'améliorer la pertinence des recommandations en comblant des lacunes de chaque approche et en prenant en considération tous les facteurs qui influenceraient les intérêts du consommateur. Ainsi, même si les approches de recommandation classiques ont déjà prouvé leur utilité et leur efficacité, la majorité n'utilise pas toutes les connaissances sur les comportements des consommateurs et sont incapables d'expliquer leurs recommandations à l'utilisateur.

Dans ce travail, nous nous intéressons à la problématique d'unification des approches de recommandation afin d'intégrer et de profiter des connaissances théoriques et empiriques des recherches sur la psychologie du consommateur. La généralisation des idées directrices des approches existantes permettrait de proposer un modèle capable de prédire les choix des consommateurs qu'ils soient influencés par les caractéristiques des produits, par leurs démographies ou par le contexte. Notre objectif se présente donc, d'une part, du point de vue méthodologique par la modélisation des comportements des consommateurs et des facteurs desquels dépendent leurs intérêts et décisions. D'autre part, nous formalisons statistiquement le modèle proposé afin de prendre en considération la nature prédictive de la problématique de recommandation et mieux gérer les incertitudes concernant les comportements observés.

3 Formalisation du modèle de recommandation proposé

Les recherches sur la psychologie du consommateur identifient plusieurs facteurs qui influencent les comportements des consommateurs (Hawkins *et al.*, 2003). Cependant, seuls les descripteurs pouvant être recueillis sur les plateformes de commerce électronique avec une interaction minimale avec les utilisateurs, on été retenus dans ce travail. Ces facteurs sont modélisés par les variables suivantes :

– Les consommateurs : chaque utilisateur $u \in U$ est décrit par un ensemble de variables démographiques d_i comme l'âge, le sexe, le niveau d'études, les revenus annuels ou le pays tel que $u = (d_1, \ldots, d_{N_d})$. De plus, les utilisateurs sont aussi décrits par l'ensemble des objectifs qu'ils tentent de satisfaire moyennant les acquisitions à effectuer similairement au travail de Zhang *et al.* (2007).

- Les ressources : chaque ressource (p.ex. produit) $x \in X$ est décrite par un ensemble d'attributs f_i tel que $x = \left(f_1, \ldots, f_{N_f}\right)$. Ces attributs permettent le calcul des similarités existantes entre les produits moyennant une mesure de similarité. Les produits peuvent aussi être décrits par les objectifs auxquels ils répondent à travers leurs propriétés (Zhang *et al.*, 2007).
- Les évaluations : les évaluations $e \in \{e_1, e_2, \ldots, e_{N_e}\}$ sont les notes que peuvent attribuer les consommateurs aux ressources. Une note e_{ij} reflète l'attitude et l'intérêt d'un utilisateur u_i vis-à-vis des attributs d'un produit x_j et leur concordance avec ses intention et ses objectifs.
- Le contexte : la prise en compte du contexte $q \in \left\{q_1, \ldots, q_{N_q}\right\}$ lors de l'analyse et la prédiction des comportements des consommateurs permet de capturer les intérêts et les achats périodiques ou occasionnels.
- L'historique des achats : l'achat est une variable binaire $a \in \{a^+, a^-\}$ et est la cible de notre modèle de recommandation. Contrairement à des approches recommandation existantes qui recommandent les items ayant la meilleure note prédite, nous estimons que la probabilité d'achat est la plus appropriée surtout dans les cas où l'acquisition a un certain coût. En effet, l'intérêt n'est pas le seul facteur qui influence les achats, mais aussi la concordance des items aux contexte et aux contraintes du consommateur.

Afin d'exploiter les similarités entre les consommateurs et les ressources, nous procédons à leur segmentation en un ensemble de classes homogènes et introduisons, deux variables représentant les catégories de produits et les groupes d'utilisateurs similaires. Le regroupement des utilisateurs similaires permet au modèle d'utiliser les opinions et les expériences du groupe pour mieux cibler un individu. De même, le recours à la segmentation des produits permet de prédire les intérêts d'un consommateur pour un produit en se basant sur ses intérêts exprimés à l'égard de ceux qui lui sont similaires. Ceci permettrait au modèle de recommander les nouveaux produits qui n'ont pas encore été vus, évalués ou achetés.

Pour simplifier le calcul des probabilités dans le modèle proposé, nous adoptons les hypothèses de dépendance et d'indépendance conditionnelles suivantes entre les variables :

- Les ressources ne dépendent que des catégories auxquelles ils appartiennent.
- Les utilisateurs ne dépendent que des groupes auxquels ils appartiennent.
- Les évaluations ne sont conditionnées que par les groupes d'utilisateurs, les catégories des ressources et du contexte. En effet, l'intérêt d'un consommateur u du groupe G par rapport à un produit x de la catégorie C pourrait être déduit à partir des évaluations attribuées par les individus du groupe G similaires à u aux produits de la catégorie C similaires à x.
- L'acte d'achat d'un produit x dépend de l'intérêt que sa catégorie C a suscité chez le groupe de consommateurs G auquel appartient l'individu considéré. Cependant, nous considérons que l'achat n'est pas une conséquence directe de l'intérêt puisqu'il dépend aussi du contexte du consommateur.

Pour pouvoir prédire les achats en utilisant le modèle proposé, nous nous intéressons à la probabilité qu'un consommateur donné achète un item donné dans un contexte précis. Cette probabilité s'exprime comme étant la probabilité d'achat sachant l'utilisateur, le produit et le contexte. En utilisant les hypothèses de dépendances et d'indépendances conditionnelles entre les variables du modèle, la probabilité recherchée s'écrit comme suit :

$$p(a^+|u,x,q) = \sum_{i=1}^{N_C} \sum_{j=1}^{N_G} \sum_{k=1}^{N_E} p(a^+|e_k,q)p(e_k|c_i,g_j,q)p(c_i|x)p(g_j|u) \quad (1)$$

$p(c_i|x)$ (resp. $p(g_j|u)$) représente le degré d'appartenance d'un item x (resp. un utilisateur u) à la catégorie c_i (resp. au groupe g_j). $p(c_i|x)$ et $p(g_j|u)$ sont reliés à l'étape de segmentation où un individu (utilisateur ou item) peut être affecté à un ou plusieurs "clusters" (groupes ou catégories). Ici, nous avons eu recours aux techniques de classification déterministe telles que *k-means* et probabiliste tel que *c-means*. Dans la classification déterministe, chaque individu est affecté exclusivement à une seule classe alors que dans une classification probabiliste, il pourrait être affecté à plusieurs classes en même temps avec des probabilités d'appartenance différentes. Cependant, nos premières expérimentations avec l'approche probabiliste ont montré de faibles résultats et une difficulté à séparer les classes achat et non-achat. Par contre, le recours à la classification déterministe permet de simplifier la probabilité d'achat puisque les probabilités $p(c_i|x)$ et $p(g_j|u)$ prennent seulement les valeurs un et zéro selon l'appartenance ou non de l'observation à la classe. Dans ce cas, la probabilité est la suivante :

$$p(a^+|u,x,q) = \sum_{k=1}^{N_E} p(a^+|e_k,q)p(e_k|c_n,g_m,q) \tag{2}$$

Le terme $p(e_k|c_i,g_j,q)$ dénote la probabilité d'observer un consommateur du groupe g_j attribuer une note e_k à un item de la catégorie c_i dans le contexte q. Ce terme fait le lien entre notre modèle et les approches de recommandation existantes. En effet, selon les données disponibles concernant l'utilisateur et le produit considérés, le degré d'intérêt prédit par ce terme pourrait être issu d'un filtrage collaboratif, par contenu ou contextuel. Dans les cas où les consommateurs n'ont pas accès à un système de notation, la valeur e_k pourrait être définie comme étant une attitude exprimée par le consommateur par rapport à un produit (p.ex. ajout aux favoris, nombre de consultations, durée de consultation). Lorsqu'elle est exprimée avant l'achat, l'attitude refléterait un certain degré d'intérêt que suscite le produit chez le consommateur. Cependant, lorsque l'attitude est exprimée suite à l'achat du produit, elle est interprétée comme étant une satisfaction ou un remords.

Le terme $p(a^+|e_k,q)$ représente la probabilité d'achat sachant l'intérêt ou la note e_k dans le contexte courant q. En effet, nous estimons que les items potentiels à la recommandation ne sont pas seulement ceux qui pourraient intéresser l'utilisateur (par rapport à la note prédite) mais aussi ceux qu'il pourrait acheter dans un contexte donné. Cette hypothèse est motivée le fait que l'achat ne dépend pas seulement de l'intérêt porté à l'item, mais aussi du contexte courant.

4 Évaluation comparative et résultats expérimentaux

Dans cette section, les détails de l'implantation du modèle de recommandation proposé sont d'abord présentés. Ensuite, les différents résultats expérimentaux sont discutés afin d'évaluer la capacité du modèle à prédire les intérêts et les achats des consommateurs et ainsi mesurer la qualité des recommandations générées.

4.1 Jeu de données

Afin d'évaluer les performances de notre proposition, nous utilisons le jeu de données de MovieLens [1] destiné à la recommandation de film puisqu'il permet d'intégrer la majorité des

1. http ://grouplens.org/datasets/movielens/

variables prévues par notre modèle. D'une part, les comportements des consommateurs sont décrits par les notes qu'ils donnent aux films qu'ils ont vus. D'autre part, le jeu de données comporte un ensemble de variables démographiques décrivant les utilisateurs ainsi qu'un ensemble de descripteurs énumérant les caractéristiques des films. Le jeu de données de MovieLens regroupe les informations suivantes :

- 1700 utilisateurs décrits par leurs âges, sexes et occupations.
- 950 films ayant chacun un identifiant, un titre, une date de sortie et un ensemble de genres parmi les 19 prédéfinis (Action, Aventure, Animation, etc...).
- 100000 notes, représentant chacune l'évaluation donnée par un utilisateur à un film ($1 \leq e_k \leq 5$). Chaque utilisateur du jeu de données a évalué au moins 20 films.

4.2 Algorithmes implémentés

Dans ce travail, quatre approches de recommandation ont été implémentées afin d'évaluer les prédictions de notre modèle. Dans ce cadre, l'objectif est de prédire la note r_{ac} que l'utilisateur actif u_a donnerait à un produit x_c, en se basant sur l'ensemble des utilisateurs $u \in U$ ainsi que sur leurs notes r_{ux} qu'ils auraient donné aux produits disponibles $x \in X$. Plusieurs variantes de chaque approche de recommandation ont été évaluées, chacune adoptant différentes implémentations de ses techniques sous-jacentes telles que les mesures de similarité (p. ex. le cosinus, le cosinus ajusté) et de distance (e.g. distance euclidienne et de Manhattan). Enfin, pour chaque approche, plusieurs techniques de segmentation (p. ex. déterministes et floues) et estimateurs de note ont été évalués afin de déterminer ceux qui sont les plus performants.

4.2.1 Filtrage collaboratif orienté utilisateur (UC-CF)

Afin d'établir une liste de produits recommandés pour un utilisateur actif u_a, cette approche commence par la sélection des utilisateurs $u_i \in S_{u_a}$ ayant les mêmes avis que celui-ci en comparant les notes qu'ils avaient donné au mêmes produits. Ensuite, la note r_{ac} que pourrait donner l'utilisateur courant u_a à un produit candidat x_c, est déduite par l'agrégation des notes r_{ic} affectées à x_c par le voisinage d'utilisateurs similaires S_{u_a} (Resnick *et al.*, 1994).

Les estimateurs suivants ont été implémentées et intégrées aux variantes évaluées de UC-CF pour l'estimation de la note r_{ac} que pourrait données un utilisateur u_a à un produit x_c :

- La moyenne : pour un utilisateur donné, l'estimation de la note qu'il pourrait donner à un produit x_c est la moyenne de celles qui lui ont été affectées par les utilisateurs qui lui sont similaires (i.e. $u_i \in S_{u_a}$).
- Moyenne pondérée : les notes sont pondérées proportionnellement au degré de similarité de leurs auteurs à l'utilisateur actif.

4.2.2 Filtrage collaboratif orienté item (IC-CF)

Pour prédire la note que donnerait l'utilisateur actif u_a à un produit x_c, l'approche procède par aggrégation des notes que u_a aurait affectée à des produits qui sont notés similairement à celui-ci. Pour celà, l'algorithme calcule $ISim(x_c, x_i) = ISim(\vec{r_c}, \vec{r_i})$, les similarités entre le produit candidat x_c et tout autre produit x_i en se basant sur les notes $\vec{r_c}$ et $\vec{r_i}$ que ces derniers ont eu de la

part des mêmes utilisateurs. Ensuite, la note estimée r_{ac} est calculée par l'agrégation des notes que u_a aurait données aux produits $x_i \in S_{x_c}$ les plus similaires à x_c (Sarwar *et al.*, 2001).

Les estimateurs de notes suivants ont été implémentés pour les différentes variantes du filtrage collaboratif orienté item :

- La moyenne : la note estimée est la moyenne des notes que u_a aurait affectées aux produits du voisinage de x_c.
- Moyenne pondérée : les notes affectées par u_a aux produits $x_i \in S_{x_c}$ sont pondérées proportionnellement à leurs similarité $ISIM(x_c, x_i)$ au produit candidat x_c.

4.2.3 Filtrage basé sur le contenu (CBF)

Cette approche part du calcul des similarité $ISim_{x_c,x_i}$ entre le produit candidat x_c et les autres x_i en se basant sur leurs descripteurs $\overrightarrow{f}_i = \{f_{i,1}, f_{i,2}, f_{i,3}, \dots\}$. Ensuite, la note prédite r_{ac} est estimée similairement au filtrage collaboratif orienté item en utilisant la moyenne et la moyenne pondérée comme estimateurs (Pazzani, 1999).

4.2.4 Filtrage démographique (DF)

Le filtrage démographique implémenté utilise à la fois les similarités entre les utilisateurs et les items afin de prédire les notes. L'approche commence par la sélection des voisinages d'utilisateurs S_{u_a} démographiquement similaires à u_i et des ressources S_{x_c} similaires à x_c sur la base de leurs attributs (Pazzani, 1999). Ensuite, la note est estimée par l'agrégation des notes attribuées par les membres de S_{u_a} aux éléments de S_{x_c} moyennant un des estimateurs suivants :

- Moyenne : si N est le nombre des notes $r_{i,j}$ tel que $u_i \in S_{u_a}$ et $x_j \in S_{x_c}$, la note estimée est calculée comme suit :

$$\hat{r}_{ac} = \frac{1}{N} \sum_{u_i \in S_{u_a}} \sum_{x_j \in S_{x_c}} r_{i,j} \tag{3}$$

- Moyenne pondérée : les notes sont pondérées par $USim(u_a, u_i)$, la similarité de l'utilisateur u_i à l'utilisateur actif u_a et/ou $ISim(x_c, x_j)$, la similarité d'un produit x_j par rapport au candidat x_c. l'estimateur est formalisé comme suit :

$$\hat{r}_{ac} = \frac{\sum_{u_i \in S_{u_a}} \sum_{x_j \in S_{x_c}} USim(u_a, u_i) \cdot ISim(x_c, x_j) \cdot r_{i,j}}{\sum_{u_i \in S_{u_a}} \sum_{x_j \in S_{x_c}} USim(u_a, u_i) \cdot ISim(x_c, x_j)} \tag{4}$$

4.2.5 Modèle de recommandation fréquentiste proposé (FM)

Afin d'adapter le modèle de recommandation proposé au jeu de données utilisé, les variables contexte et achat non fournies sont omises. Par conséquent, l'objectif du modèle adapté est de calculer la probabilité qu'une note de valeur e_k soit affectée par l'utilisateur u à un produit x afin de déterminer son intérêt à celui-ci. Cette probabilité s'écrit comme suit :

$$p(e_k|u,x,q) = p(e_k|g_u, c_x, q)p(g_u|u)p(c_x|x) \tag{5}$$

Afin de générer des recommandations, l'algorithme commence par la segmentation des utilisateurs et des produits. Dans ce travail, l'étape de segmentation a été effectuée par les algorithmes de Kmeans et de maximisation de l'espérance en utilisant les distances euclidienne et de

Manhattan. Pour cela, plusieurs variables ont été recodées en variables binaires tel que l'occupation des utilisateurs et les genres des films. Les premières expérimentations montrent que ces mesures génèrent des classifications similaires et donc des résultats de prédiction très proches. Cependant, les meilleurs résultats ont été obtenus en utilisant la distance euclidienne puisqu'elle prends en considération le caractère ordinal des variables ordinales (p.ex. l'âge de l'utilisateur et la date de sortie d'un film). Par ailleurs, plusieurs méthodologies ont été employées.

1. Segmentation par les caractéristiques intrinsèques : la segmentation se base sur les attributs démographiques pour les utilisateurs et sur les caractéristiques pour les produits.
2. Segmentation par les notes : la segmentation des utilisateurs est effectuée sur la base de la similarité de leurs notes. Cette approche est similaire à celle du filtrage collaboratif orienté utilisateur. Analogiquement, les produits sont segmentés sur la base de la similarité des notes qu'ils ont obtenues de la part des mêmes utilisateurs. Cette approche est analogique à celle utilisée dans le filtrage collaboratif orienté item.
3. Segmentation mixte : la segmentation se base à la fois sur les caractéristiques intrinsèques et les notes. Ceci permet de rassembler dans les mêmes groupes les utilisateurs appartenant aux mêmes classes démographiques et ayant les mêmes centres d'intérêt. De même, les catégories de produits rassembleraient des produits ayant des caractéristiques similaires et/ou appréciées par les mêmes utilisateurs.

Afin d'estimer la note $\hat{r}_{ux}$ affectée par un utilisateur u à un produit x, la probabilité $p(e_k|u,x,q)$ est utilisée pour calculer l'espérance de la variable note tel que :

$$\hat{r}_{ux} = E[p(e|u,x,q)] = \sum_k e_k \times p(e_k|u,x,q) \tag{6}$$

4.3 Principaux résultats expérimentaux

La figure 1 présente la distribution des erreurs de prédiction des notes des utilisateurs pour les meilleures variantes des approches étudiées. Les valeurs d'erreur obtenues ainsi sont résumées dans des boites à moustache afin d'étudier la variabilité de la qualité de chacune des approches. Les boites représentent ainsi les valeurs d'erreur de prédiction obtenues entre le premier et le troisième quartile, tandis que la ligne interne représente la valeur médiane de l'erreur. Les valeurs d'erreur extrêmes sont représentées par des points à l'extérieur des boites.

Le tableau 1 présente les résultats expérimentaux obtenus par validation croisée (2 échantillonnages), pour chacune des approches étudiées en termes de précision, de rappel, d'erreur absolue moyenne (MAE) et de racine de l'erreur quadratique moyenne (RMSE). Les mesures de précision et de rappel sont calculées à partir des notes estimées par chaque approche, en considérant chacune des valeurs possibles de cette variable (entre 1 et 5) comme étant une classe à prédire. Le tableau 1 montre que les meilleurs résultats sont obtenus par le modèle proposé ainsi que par les approches de filtrage collaboratif.

Les approches de filtrage par contenu (CBF) et collaboratif orienté item (IC-CF) présentent les résultats les plus proches des notres par rapport à l'erreur absolue moyenne. Cependant, leur pertinence en termes de rappel et de précision de ces approches varie largement en fonction de la taille du voisinage de similarité utilisée lors de la détermination des produits similaires au produit candidat à la recommandation.

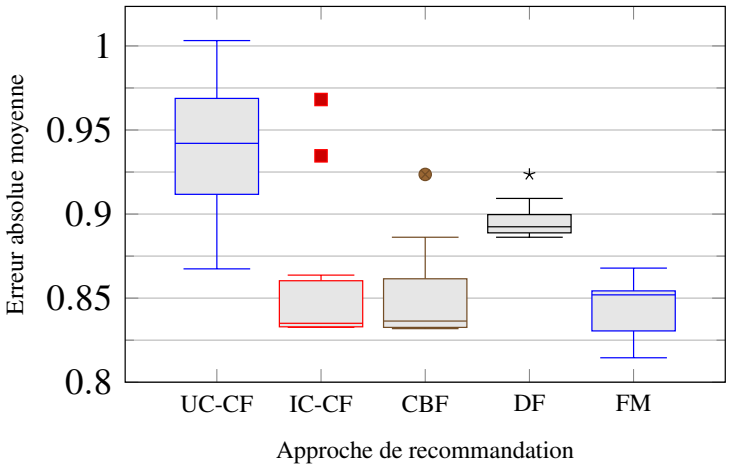

FIGURE 1 – Résultats comparatifs (erreurs de prédiction)

TABLE 1 – Résultats expérimentaux

	DF	CBF	IC-CF	UC-CF	FM
Précision	20,39%	23,85%	27,71%	28,47%	28,61%
Rappel	17,07%	29,11%	44,96%	43,21%	46,39%
MAE	0,8862	0,8318	0,8325	0,8674	0,8145
RMSE	1,1225	1,0446	1,0432	1,0981	1,0539

Le filtrage basé sur le contenu (CBF) effectue ses recommandations en tirant profit des similarités entre les produits. Cependant, la mesure de similarité ainsi que l'estimateur de note employés ont moins d'influence sur la qualité des recommandations que les données utilisées. En effet, lorsque les entités du domaine d'application sont homogènes et font partie du même concept sémantique, cette approche est capable de générer des recommandations pertinentes puisqu'elle favorise les contenus similaires à ceux déjà appréciés par l'utilisateur. Cependant, lorsque les ressources disponibles appartiennent à des catégories sémantiquement différentes, cette approche est incapable de recommander des produits appartenant à des catégories inconnues par l'utilisateur. En effet, les produits de différentes catégories ne sont pas comparables à cause de leurs différents descripteurs et limitent ainsi l'applicabilité des mesures de similarité.

Les expérimentations montrent que le filtrage collaboratif orienté utilisateur (UC-CF) est le plus sensible à la configuration. Sur le jeu de données de MovieLens, les meilleurs résultats ont été obtenus par l'estimateur de note basé sur la moyenne pondérée. La qualité de cet estimateur s'explique par le recours à la pondération des notes des utilisateurs par rapport à leurs similarités. Ces pondérations donnent ainsi plus d'importance aux avis des utilisateurs qui ont plus de similarité démographique et/ou comportementale avec l'utilisateur actif.

Le filtrage démographique (DF) donne des résultats homogènes indépendamment de la mesure de similarité et de l'estimateur de note employés. Cependant, la qualité de ses prédictions dépend des tailles des voisinages d'utilisateurs et de produits utilisés qui nécessitent une phase de configuration afin de maximiser la qualité et les performances de l'approche.

Les expérimentations présentées montrent que le modèle de recommandation proposé per-

met de décrire et d'anticiper les comportements des consommateurs. En effet, en unifiant et en généralisant les idées directrices des approches de recommandation existantes, notre modèle est capable de prédire et de quantifier les intérêts des consommateurs, qu'ils soient influencés par la démographie ou par les caractéristiques des produits. Dans ce modèle, le nombre de groupes d'utilisateurs N_G et de catégories de produits N_C influencent la qualité des recommandations en termes de rappel et de précision. En effet, sous-estimer N_g ou N_c peut conduire à une perte de précision, puisque les consommateurs (ou les produits) au sein du même groupe deviennent moins similaires. De même, lorsque le nombre des classes est surestimé, le rappel du modèle peut diminuer puisque les individus similaires (produits ou clients) peuvent être affectés à différents groupes et ne serait donc pas pris en compte lors de la recommandation. Dans les expérimentations présentées, N_g ou N_c ont été déterminés de manière empirique guidée par la visualisation des distributions des utilisateurs et des films. Par ailleurs, le seuil de probabilité au-dessus duquel un produit est recommandé pourrait être ajusté de manière à trouver un compromis entre l'exactitude des recommandations et leur diversité. En effet, l'augmentation de ce seuil est adaptée aux utilisateurs ayant des besoins spécifiques puisqu'elle conduit à moins de recommandations, mais avec une précision élevée. Par contre, le seuil peut être réduit pour promouvoir la diversité des recommandations pour les consommateurs impulsifs et les clients errants n'ayant pas de produits spécifiques à acheter.

5 Conclusion et perspectives

Le travail présenté dans cet article a pour objectif de proposer un modèle de recommandation unifié capable de prédire les intérêts et les achats des consommateurs. Le modèle proposé se base sur un ensemble de facteurs et d'hypothèses de dépendance afin de prédire de manière probabiliste les comportements de consommation. Les principales contributions de ce travail proviennent de (1) la modélisation des comportements de consommation, (2) de la généralisation des idées directrices des principales approches de recommandation existantes et enfin (3) de la formalisation statistique du modèle et de la méthodologie d'inférence des recommandations. L'expérimentation comparative réalisée sur des données réelles a permis de positionner notre proposition par rapport aux principales approches existantes et de valider son apport.

Comme perspective de ce travail, nous considérons l'intégration de la description visuelle des ressources dans le modèle de recommandation proposé afin de mieux couvrir les facteurs qui influenceraient les choix des consommateurs. Ceci nécessite auparavant l'étude de son influence et l'évaluation de son apport au modèle de recommandation proposé.

Références

AGRAWAL R. & SRIKANT R. (1994). Fast algorithms for mining association rules in large databases. In J. B. BOCCA, M. JARKE & C. ZANIOLO, Eds., *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile*, p. 487–499 : Morgan Kaufmann.

BAAZAOUI H., HADDAD M. R. & GHEZALA H. B. (2014). A personalised semantic and spatial information retrieval system based on user's modelling and accessibility measure. *International Journal of Multicriteria Decision Making*, **4**(2), 183–200.

BALABANOVIĆ M. & SHOHAM Y. (1997). Fab : content-based, collaborative recommendation. *Communications of the ACM*, **40**(3), 66–72.

BOUTEMEDJET S. & ZIOU D. (2008). A graphical model for context-aware visual content recommendation. *IEEE Transactions on Multimedia*, **10**(1), 52–62.

BURKE R. (2000). Knowledge-based recommender systems. In *ENCYCLOPEDIA OF LIBRARY AND INFORMATION SYSTEMS*, p. 2000 : Marcel Dekker.

BURKE R. D. (2007). Hybrid web recommender systems. In P. BRUSILOVSKY, A. KOBSA & W. NEJDL, Eds., *The Adaptive Web, Methods and Strategies of Web Personalization*, volume 4321 of *Lecture Notes in Computer Science*, p. 377–408 : Springer.

DESHPANDE M. & KARYPIS G. (2004). Item-based top-*N* recommendation algorithms. *ACM Transactions on Information Systems*, **22**(1), 143–177.

GOLDBERG D., NICHOLS D., OKI B. M. & TERRY D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, **35**(12), 61–70.

HAWKINS D. I., BEST R. J. & CONEY K. A. (2003). *Consumer Behavior : Building Marketing Strategy*. McGraw-Hill/Irwin, 9 edition.

HSIEN LIAO S., HUI CHU P., JU CHEN Y. & CHANG C.-C. (2012). Mining customer knowledge for exploring online group buying behavior. *Expert Systems with Applications*, **39**(3), 3708 – 3716.

IYENGAR S. S. & LEPPER M. R. (2000). When choice is demotivating : can one desire too much of a good thing ? *Journal of Personality and Social Psychology*, **79**(6), 995–1006.

JACOBY J., SPELLER D. E. & BERNING C. A. K. (1974). Brand choice behavior as a function of information load : Replication and extension. *Journal of Consumer Research : An Interdisciplinary Quarterly*, **1**(1), 33–42.

JONES G. J. F. (2005). Challenges and opportunities of context-aware information access. In *UDM*, p. 53–62 : IEEE Computer Society.

KRULWICH B. (1997). Lifestyle finder : Intelligent user profiling using large-scale demographic data. *AI Magazine*, **18**(2), 37–45.

LIN W., ALVAREZ S. A. & RUIZ C. (2002). Efficient adaptive-support association rule mining for recommender systems. *Data Min. Knowl. Discov*, **6**(1), 83–105.

LINDEN G., SMITH B. & YORK J. (2003). Amazon.com recommendations : Item-to-item collaborative filtering. *IEEE Internet Computing*, **7**(1), 76–80.

MOONEY R. J. & ROY L. (2000). Content-based book recommending using learning for text categorization. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, p. 195–204. New York : ACM Press.

PAZZANI M. J. (1999). A framework for collaborative, content-based and demographic filtering. *Artif. Intell. Rev*, **13**(5-6), 393–408.

POUSSEVIN M., GUARDIA-SEBAOUN E., GUIGUE V. & GALLINARI P. (2014). Recommandation par combinaison de filtrage collaboratif et d'analyse de sentiments. In *CORIA-CIFED*, p. 27–42.

RESNICK P., IACOVOU N., SUCHAK M., BERGSTORM P. & RIEDL J. (1994). Grouplens : An open architecture for collaborative filtering of netnews. In *Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work*, p. 175–186 : ACM.

SARWAR B. M., KARYPIS G., KONSTAN J. A. & RIEDL J. (2001). Item-based collaborative filtering recommendation algorithms. In *WWW*, p. 285–295.

SCHWARTZ B. (2005). *The paradox of choice : Why more is less*. Harper Perennial.

WOERNDL W. & GROH G. (2007). Utilizing physical and social context to improve recommender systems. In *Web Intelligence/IAT Workshops*, p. 123–128 : IEEE.

ZHANG Y., QI J., SHU H. & CAO J. (2007). Personalized product recommendation based on customer value hierarchy. In *SMC*, p. 3250–3254 : IEEE.