Utilité et perception de la diversité dans les systèmes de recommandation

Sylvain Castagnos, Armelle Brun, and Anne Boyer

Équipe KIWI - LORIA
Campus Scientifique, B.P. 239
54506 Vandœuvre - France
{Sylvain.Castagnos, Armelle.Brun, Anne.Boyer}@loria.fr

RÉSUMÉ. *De récentes études ont montré que la diversité dans les systèmes de recommandation est positivement corrélée à la satisfaction des utilisateurs et renforce/facilite leur choix d'un item (Castagnos* et al., *2010). Si l'impact de cette nouvelle dimension a été mesuré, les raisons d'un tel succès restent cependant encore inexpliquées. Forts de ce constat, notre objectif est d'analyser plus finement l'utilité réelle et perçue de la diversité dans les systèmes de recommandation. Dans cette optique, nous avons réalisé une étude auprès de 250 utilisateurs permettant de comparer 5 approches (mêlant filtrage collaboratif, filtrage par contenu et popularité) avec différents degrés de diversité. Les résultats montrent que la diversité dans les recommandations est perçue par les utilisateurs et améliore leur satisfaction, même si elle suscite parfois méfiance ou incompréhension. En outre, cette étude a mis en lumière la nécessité de constituer des modèles de préférences suffisamment divers pour générer de bonnes recommandations.*

ABSTRACT. *Recent studies have highlighted the correlation between the degree of diversity within recommender systems and users' satisfaction, and the fact that diversity increases confidence within the choice of an item (Castagnos* et al., *2010). We now need to understand the reasons of the positive impact of diversity on recommender systems. We thus decided to design a user study focused on the utility and perceived qualities of this new dimension. We recruited 250 users and compared 5 different approaches based on collaborative filtering, content-based filtering and popularity, with different degrees of diversity. Results show that diversity, when recommendations are made explicit, may reduce users' acceptance rate. However, it helps increasing users' satisfaction. Besides, this study has highlighted the need to build preference models that are diverse enough, so as to generate good recommendations.*

MOTS-CLÉS : *Systèmes de recommandation, Étude utilisateurs, Diversité, Modélisation.*

KEYWORDS: *Recommender Systems, User Study, Diversity, User Modeling.*

1. Introduction

Les systèmes de recommandation ont pour but de faciliter la recherche et l'accès à l'information en proposant des items adaptés aux besoins et préférences des utilisateurs. Deux décennies de recherche fondamentale et de transferts technologiques dans de nombreux domaines applicatifs ont permis d'illustrer la valeur ajoutée et l'utilité de tels systèmes. Néanmoins, paradoxe dans la quête de la personnalisation, la plupart des modèles élaborés jusqu'ici s'attachent à maximiser la précision des recommandations sans jamais intégrer les facteurs humains inhérents au processus de décision. A titre d'exemple, le concours Netflix [1] se concluait en 2009 sur la victoire de l'équipe BellKor's Pragmatic Chaos, après trois années d'efforts pour améliorer l'erreur quadratique moyenne (RMSE) des meilleurs algorithmes de seulement deux centièmes (Koren *et al.,* 2009, Sill *et al.,* 2009, Töscher *et al.,* 2009). Ironie du sort, l'algorithme ayant remporté le concours ne sera jamais utilisé, rendu obsolète par de nouveaux modes d'interaction et l'observation de nouveaux comportements chez les utilisateurs (Sawers, 2012).

Pratiquement au même moment, des travaux visant à mesurer l'acceptation et l'adoption des systèmes de recommandation par le grand public montraient qu'une différence de qualité pouvant aller jusqu'à 10% entre deux algorithmes n'était pas perçue par les utilisateurs (Castagnos *et al.,* 2010, Jones, 2010). En revanche, ces mêmes études faisaient ressortir l'influence de différents facteurs humains sur la satisfaction des utilisateurs, tels que la confiance acquise grâce à différents modes de regroupement des recommandations et aux explications fournies, ou encore le besoin en diversité dans les recommandations. Ce dernier résultat, relatif à la diversité, constituera le point de départ de notre travail. En effet, les expériences menées en 2010 par (Castagnos *et al.,* 2010) avaient pour objectif initial d'identifier les différentes sous-étapes inhérentes au processus de décision lorsqu'un utilisateur est confronté à un système de recommandation. Ce facteur de diversité n'avait pas été anticipé. Il est donc rapidement apparu comme une dimension clef dans le processus de décision, sans que l'expérience ne permette d'en cerner tous les tenants et aboutissants.

En résumé, dans le cadre de ce papier, nous avons réalisé une nouvelle étude utilisateur avec le dessein de mieux comprendre le rôle et l'impact de la diversité dans les systèmes de recommandation. Il s'est agi dans un premier temps d'implémenter plusieurs familles d'algorithmes, respectivement basées sur le filtrage collaboratif (**FC**), le filtrage par contenu (**FBC**) et le filtrage par popularité (**POP**). Une deuxième contribution de ce papier a consisté à proposer deux nouveaux algorithmes hybrides combinant filtrage collaboratif et filtrage par contenu, appelés **FCDR** et **FCDF**, permettant de réguler le degré de diversité des recommandations. La comparaison de ces algorithmes a ensuite nécessité l'emploi d'un modèle inter-groupes à 5 niveaux (1 groupe pour chaque algorithme) et d'un modèle intra-sujet à 2 niveaux (recommandations fournies de façon implicite et explicite) dans le domaine du cinéma. La base d'apprentissage nécessaire à cette expérimentation comportait la description complète

1. http://www.netflixprize.com/

de plus de 500 films et les votes de 3000 utilisateurs. Cette étude a été réalisée dans un intervalle de temps d'une semaine et a permis de collecter les données de 250 utilisateurs, répartis aléatoirement en groupes de 50 personnes.

Les résultats de cette étude ont permis, d'une part, de confirmer l'impact positif de la diversité sur la satisfaction des utilisateurs. D'autre part et de manière plus surprenante, ils ont révélé l'importance de constituer des modèles de préférences suffisamment divers, notamment pendant la phase de démarrage à froid, en incitant les utilisateurs à voter pour des items différents. Enfin ils ont montré que, malgré son caractère bénéfique, la diversité doit être utilisée avec parcimonie car elle peut susciter la méfiance et l'incompréhension de certains utilisateurs.

La suite de ce papier se décompose comme suit : la partie 2 présentera un état de l'art de la diversité dans les systèmes de recommandation, à la fois du point de vue conception que du point de vue évaluation des systèmes. La section 3 sera dédiée à la description de notre expérimentation. La partie 4 sera consacrée à la présentation et à la discussion des résultats. Enfin, nous conclurons ce papier.

2. État de l'art

La diversité est un thème de recherche en plein essor dans le domaine des systèmes de recommandation. Il est couramment admis qu'elle joue un rôle dans l'amélioration des interactions entre les systèmes de recherche et d'accès à l'information et les usagers (McGinty *et al.,* 2003, Castagnos *et al.,* 2009). Néanmoins, la question de savoir **pourquoi** et **comment** améliorer la diversité reste une problématique de recherche ouverte. C'est ainsi que l'on distingue dans la littérature deux types de travaux : ceux s'attachant à mieux appréhender le rôle et l'impact de la diversité dans la prise de décision en analysant le comportement des utilisateurs, et ceux visant à intégrer cette nouvelle dimension dans les méthodes d'apprentissage automatique couramment utilisés dans les systèmes de recommandation. Les deux sous-sections suivantes présenteront un état de l'art de ces différentes investigations. Ce papier se situe à la frontière entre ces deux problématiques.

2.1. *Rôle et impact de la diversité*

La notion de diversité a été définie par (Smyth *et al.,* 2001) comme étant la dimension opposée à la similarité dans les systèmes de recommandation. Nous compléterons ici cette définition en présentant la diversité comme une mesure quantifiant la dissimilarité dans un ensemble d'items. On comprend alors qu'introduire de la diversité consiste à trouver le meilleur compromis (Adomavicius *et al.,* 2008) pour suggérer à l'utilisateur des items suffisamment similaires à ses préférences connues, sans pour autant : **(1)** scléroser les recommandations (*i.e.* ne jamais proposer de nouveauté), ni **(2)** proposer des items trop proches les uns des autres. Dans le premier cas, on parle de diversité intrinsèque permettant d'éviter la redondance dans les items suggérés (Clarke

et al., 2008). Dans le deuxième cas, il s'agit de diversité extrinsèque, dont le but est de pallier l'incertitude liée à l'ambiguité ou aux données manquantes dans les modèles de préférences utilisateurs en proposant un panel assez large de recommandations (Radlinski *et al.,* 2009). Dans les deux cas, les mécanismes d'introduction de diversité reposent sur les mêmes métriques (cf. section 2.2).

Notons également qu'une nouvelle classification de la diversité a été proposée en 2012 par (Adomavicius *et al.,* 2012), distinguant la diversité individuelle de la diversité aggrégée selon que l'on s'intéresse aux recommandations pour un individu ou pour un groupe. Dans le cadre de ce papier, nous nous focaliserons sur la notion de diversité individuelle.

La première étude à véritablement s'intéresser au rôle de la diversité fut menée dans le domaine des systèmes de recommandation conversationnels (McGinty *et al.,* 2003). Cette expérimentation a prouvé pour la première fois que la diversité a la capacité d'améliorer significativement l'efficacité des recommandations. Par la suite, (Zhang *et al.,* 2008) et (Lathia, 2010) iront jusqu'à parler de frustration de l'utilisateur en l'absence de diversité. L'étude de (McGinty *et al.,* 2003) a également mis en lumière les nombreux enjeux liés à cette nouvelle dimension. L'un de ces défis relève du fait que la diversité n'est pas préconisée à chaque cycle de recommandation.

Pour bien comprendre ce point, il faut examiner les travaux de (Häubl *et al.,* 2003) sur l'étude des différentes étapes qui jalonnent le processus de sélection d'un item dans les systèmes de recherche. Ce processus se déroule en deux temps. Lors de la première phase, l'utilisateur utilise les moyens mis à sa disposition via l'interface du système pour identifier les critères pertinents dans le cadre de sa recherche. L'identification de ces critères spécifiques à la recherche en cours se font le plus souvent par un jeu d'essais-erreurs que l'on appelle cycles de recommandations. Une fois les critères adéquats identifiés, la deuxième phase consistant à comparer toutes les caractéristiques d'un sous-ensemble d'alternatives valables peut commencer. Le champ d'étude de (Häubl *et al.,* 2003) se limitait aux domaines des systèmes de recherche et des assistants intelligents, où l'utilisateur doit effectuer une démarche volontaire dans sa recherche d'information. Néanmoins, des schémas comportementaux similaires ont été observés lors des interactions avec les systèmes de recommandation (Castagnos *et al.,* 2009). L'utilisateur a inconsciemment recours à ce même jeu d'essais-erreurs consistant à cliquer sur un item, à survoler rapidement les recommandations et à revenir le cas échéant en arrière. Une fois qu'il trouve un point de départ satisfait, il prolonge plus facilement son exploration via les recommandations et sur une plus grande profondeur (*i.e.* consultation des recommandations, ainsi que des recommandations en lien avec les recommandations cliquées, etc.).

Une fois ce modèle comportemental admis, on pressent comme l'avaient souligné les auteurs de (McGinty *et al.,* 2003) que la diversité n'est pas systématiquement souhaitable dans les cycles de recommandation, alors même que l'utilisateur cherche à renforcer sa confiance dans le système. Cette intuition a été confirmée dans (Castagnos *et al.,* 2010), à travers une étude utilisateur visant à mesurer l'évolution du besoin en diversité au cours du temps et à proposer un modèle dynamique de la diversité.

(Lathia, 2010) inscrira également la dimension temporelle dans ses travaux autour de la diversité.

De nombreux autres débats ont émergé ces dernières années sur le rôle de la diversité. A titre d'exemple, McNee *et al.* ont étudié les limites des métriques de précision employés dans les systèmes de recommandation, et ont montré que des recommandations moins précises ne sont pas nécessairement moins pertinentes (McNee *et al.*, 2006). Ils se sont également intéressés à la différence de diversité proposée entre des utilisateurs qui reviennent régulièrement sur un site et ceux qui visitent ce site pour la première fois.

Plusieurs auteurs soulignent également que le filtrage collaboratif fournit une certaine forme de diversité, en comparaison des algorithmes basés sur le contenu, via la sérendipité (*i.e.* la découverte inattendue grâce au hasard) (Herlocker *et al.*, 2004, Barragáns-Martínez *et al.*, 2010).

Forts de ces différentes conclusions sur l'attrait de la diversité, certains chercheurs ont souhaité comprendre comment présenter et/ou expliquer au mieux les items pour que la diversité soit perçue par les utilisateurs (Yu *et al.*, 2009, Ge *et al.*, 2011, Hu *et al.*, 2011).

D'autres, enfin, se sont attelés à l'intégration de cette dimension dans les modèles d'apprentissage automatique utilisés dans les systèmes de recommandation. La sous-section suivante présentera un état de l'art de solutions existantes.

2.2. *Intégration de la diversité dans les systèmes de recommandation*

Le fonctionnement d'un système de recommandation, quel qu'il soit, se décompose en trois étapes (Castagnos, 2008) : (1) la collecte implicite ou explicite des traces laissées par l'utilisateur courant à travers ses interactions avec le système (préférences, goûts, usages et contexte); (2) la synthèse de ces traces d'interaction sous forme d'un modèle utilisateur (parfois plus simplement appelé modèle de préférences selon l'approche utilisée); (3) l'exploitation de ce modèle et l'emploi de méthodes d'apprentissage automatique (*i.e.* familles d'algorithmes) pour déterminer les recommandations adaptées à l'utilisateur courant.

Les méthodes d'apprentissage sont traditionnellement classées en deux familles d'algorithmes (Ricci *et al.*, 2011) : le filtrage collaboratif et le filtrage par contenu. Jusqu'à très récemment, les mécanismes d'amélioration de la diversité étaient pour la plupart adaptés aux algorithmes de filtrage par contenu (Bradley *et al.*, 2001, McSherry, 2002, Agrawal *et al.*, 2009). Ces mécanismes peuvent intervenir directement au niveau de la métrique, et/ou au niveau de l'algorithme de classement utilisé pour les recommandations. L'objectif de ces approches est d'accroître la diversification au niveau des attributs des items. A titre d'exemple, le prix, la résolution, la marque et l'optique constitutent un ensemble d'attributs d'un appareil photographique. Il s'agit

ensuite de proposer des recommandations adaptées ne prenant pas les mêmes valeurs d'attributs.

Dans (Smyth *et al.*, 2001), la diversité est caractérisée par plusieurs métriques reposant directement sur la similarité entre items : plus les items sont similaires et homogènes, moins il y a de diversité entre eux. Ils commencent donc par définir la similarité entre deux items i_1 et i_2 comme une moyenne pondérée des similarités sur les n attributs, comme explicité dans l'équation (1).

$$Similarity(i_1, i_2) = \frac{\sum_{j=1..n} w_j * sim_{attribute=j}(i_1, i_2)}{\sum_{j=1..n} w_j} \quad [1]$$

A partir de cette métrique de similarité, (Smyth *et al.*, 2001) introduisent ensuite deux nouvelles mesures de diversité. La première, sobrement appelée $Diversity$, permet de calculer la dissimilarité moyenne au sein d'une classe C composée de m items. La deuxième est une diversité relative ($RelDiversity$) permettant de mesurer la valeur ajoutée en terme de diversité d'un item par rapport à une classe d'items C (avec une cardinalité $Card(C) = m$). Cette dernière métrique est explicitée dans l'équation (2).

$$RelDiversity(i, C) = \begin{cases} 0 \text{ si C = \{\},} \\ \frac{\sum_{j=1..m}(1 - Similarity(i, c_j))}{m} \text{ sinon.} \end{cases} \quad [2]$$

Ces différentes métriques ont ensuite été utilisées dans de nombreux travaux pour réordonner la liste des recommandations en fonction du critère de diversité dans le filtrage par contenu. On distingue alors deux types d'approches : celles qui traitent ce processus de tri comme un problème de *clustering* (Wan *et al.*, 2011), et celles basées sur une méthode de sélection (Bradley *et al.*, 2001). Dans les approches de *clustering*, l'objectif est de construire des classes d'items optimales par rapport au critère de diversité (c'est-à-dire proposant la plus grande diversité possible). Les méthodes de sélection, quant à elles, évaluent les stratégies qui apportent de la diversité dans les systèmes de recommandation, sans pour autant compromettre la précision.

Bradley et Smyth ont été parmi les premiers à proposer un algorithme de sélection glouton borné pour retrouver les items les plus similaires à une requête de l'utilisateur, mais dans le même temps divers entre eux (Bradley *et al.*, 2001). Cet algorithme consiste dans un premier temps à sélectionner les K items les plus similaires à la cible t grâce à l'équation (1). Par la suite, il complète la liste de recommandations itérativement en choisissant à chaque étape l'item i de meilleure qualité (cf. équation (3)), jusqu'à obtenir les recommandations du $top - N$ ($K < N$).

$$Quality(i, t, C) = Similarity(i, t) * RelDiversity(t, C) \quad [3]$$

Les algorithmes de reclassement des résultats du $top-N$ comme ceux de (Bradley *et al.*, 2001) sont connus pour leur excellent compromis entre rapidité des calculs et

qualité des recommandations (qualité relative à la précision et à la diversité simultanément). Ainsi, Radlinski *et al.* proposent trois méthodes alternatives reposant sur la reformulation des requêtes pour accroître la diversité dans le $top - N$ (Radlinski *et al.*, 2009). Zhang et Hurley suggèrent quant à eux de maximiser la diversité tout en maintenant une similarité adéquate en traitant cette approche comme un problème d'optimisation binaire (Zhang *et al.*, 2008). Ils ont évalué leur approche sur le corpus MovieLens[2] avec un succès raisonnable (Zhang *et al.*, 2008).

Outre les algorithmes de filtrage par contenu, certains travaux se sont intéressés à la manière d'apporter la diversité dans le filtrage collaboratif. Ziegler et McNee furent des précurseurs en proposant un formalisme générique basé sur une mesure de similarité intra-liste (ILS) et la sélection du $top - N$, qui soit intégrable dans plusieurs familles d'algorithmes dont le filtrage collaboratif (Ziegler *et al.*, 2005). (Said *et al.*, 2012) ont quant à eux étudié la possibilité d'injecter de la diversité dans le filtrage collaboratif en adaptant des algorithmes de *clustering*. Soulignons que ces trois derniers travaux sont axés sur du filtrage collaboratif pur. En conséquence, ils exploitent des métriques de similarité basées sur les votes, et non sur les attributs.

2.3. *Discussion*

L'objectif premier de ce papier est de comprendre plus finement l'impact et l'utilité que peut avoir la diversité lors des interactions entre un système de recommandation et ses utilisateurs. L'état de l'art de la section 2.1 a permis d'illustrer le caractère complexe de cette dimension, sans pour autant en cerner tous les tenants et aboutissants. En effet, les études menées jusqu'ici ont consisté à mesurer l'impact de la diversité sur la satisfaction *a posteriori*, par le truchement de post-questionnaires (Jones, 2010). Par ailleurs, si de nombreuses études (McGinty *et al.*, 2003, Zhang *et al.*, 2008, Lathia, 2010) ont mesuré l'impact de la diversité sur la satisfaction vis-à-vis du filtrage par contenu, aucune étude utilisateur comparable n'a été réalisée pour le filtrage collaboratif. (Herlocker *et al.*, 2004) ont bien pressenti l'apport en diversité via la sérendipité dans cette famille d'algorithmes, mais le degré de diversité n'y étant pas contrôlé, il n'est pas toujours garanti.

Dans ce papier, nous proposons donc de réaliser une étude utilisateur conçue autour de la notion de diversité et permettant :

– de comparer différentes familles d'algorithmes (filtrage collaboratif, filtrage par contenu, filtrage par popularité) ;

– d'observer le comportement des utilisateurs depuis la collecte des traces jusqu'au choix des recommandations, pour vérifier si la diversité n'exerce un rôle qu'au niveau du calcul de ces recommandations comme le laisse penser la littérature sur le sujet ;

– d'étudier la perception de la diversité et les différences de comportements, lorsque les recommandations sont fournies de manière implicite ou explicite.

2. http://www.grouplens.org/node/73

En outre, l'état de l'art de la section 2.2 nous permet de constater l'absence d'algorithmes hybrides filtrage collaboratif/filtrage par contenu dont l'objectif serait de préserver un équilibre entre précision et diversité des recommandations. Pour pallier cette absence, nous nous sommes inspirés des travaux de (Bradley *et al.,* 2001) et (Ziegler *et al.,* 2005) pour concevoir deux algorithmes, combinant et tirant parti du filtrage collaboratif et du filtrage par attributs pour obtenir le niveau de diversité désiré.

Au final, nous proposons donc de comparer 5 algorithmes différents dans le cadre de notre étude. La section suivante fournira les détails liés à notre expérimentation.

3. Expérience

3.1. *Support*

Nous avons choisi de réaliser notre étude dans le domaine du cinéma, pour ses nombreux avantages. D'une part, il est assez facile de collecter un grand nombre de données sur des films pour mener l'expérience dans un contexte réaliste. D'autre part, les films comportent un grand nombre d'attributs et sont très souvent notés par les utilisateurs, contrairement à d'autres types d'items, ce qui en fait un cadre d'étude idéal pour les différents algorithmes que nous souhaitons comparer. Enfin, il s'agit d'un domaine assez populaire et familier pour les utilisateurs. Cela maximise les chances que les utilisateurs connaissent suffisamment d'items dans la liste proposée et puissent exprimer leurs préférences.

Pour mener à bien notre étude, nous avons mis en place un site Web. Par manque de place, nous ne pourrons pas mettre de capture d'écran dans cet article, mais le site est accessible en ligne[3].

Nous avons commencé par collecter un maximum d'informations sur les contenus de plus de 500 films. Ceci inclut les titres et résumés en anglais et en français, les images, les bandes-annonces, les votes moyens des spectateurs et de la presse (ainsi que le nombre de votants), les genres, les acteurs, les réalisateurs, les scénaristes, l'année de production, la durée et l'appartenance à une saga (ex. : Indiana Jones 1 à 4). Tous ces éléments permettent à la fois à l'utilisateur de se remémorer le contenu des différents films et au système d'appliquer des algorithmes de filtrage par contenu.

Le filtrage collaboratif, quant à lui, nécessite les votes individuels d'un grand nombre d'utilisateurs. Pour cette raison, nous avons collecté les votes de 3.158 utilisateurs. Chaque utilisateur a fourni au moins 20 votes et chaque film a été voté par au moins 20 utilisateurs. Ces seuils correspondent au nombre minimum de votes estimé par (Schickel *et al.,* 2006) pour obtenir des recommandations de bonne qualité dans un algorithme de filtrage collaboratif. Nous avons dû construire nous-mêmes cette base, plutôt que de réutiliser les corpus MovieLens ou NetFlix, afin de nous assurer

3. http://www.movit.tv/tut5/index.php

qu'il est toujours possible de calculer les similarités entre films quels que soient les attributs utilisés. Toutefois, la taille de notre corpus est tout à fait comparable à celle de MovieLens. Par ailleurs, et contrairement à MovieLens, nous nous sommes assurés d'avoir une bonne répartition des films en terme de popularité. Un seuil minimum du nombre de votants, pour un film donné, sur IMDb a été fixé à 200 et une vérification auprès d'un panel d'une vingtaine d'utilisateurs a été effectuée pour garantir que tous les films de notre base sont relativement connus. De même, nous avons choisi les films aléatoirement, tout en veillant à avoir une bonne répartition des votes sur l'échelle des valeurs possibles allant de 1 à 5 (et notamment une bonne représentativité parmi le top-250 et le bottom-100 des films IMDb). La moyenne des votes des 3.158 utilisateurs de la base d'apprentissage est égale à 3,66 avec un écart type de 1,37.

L'ensemble des informations sur les films et sur les votes constituent notre base d'apprentissage. Cette dernière a été créée grâce aux APIs d'Allociné et d'IMDb. Les caractéristiques de notre base d'apprentissage sont résumées dans le tableau 1.

Type	Nombre	Type	Nombre
Films	509	Genres	23
Acteurs	903	Pays de production	17
Réalisateurs	310	Sagas	98
Scénaristes	351	Votes	173.120

Tableau 1. *Caractéristiques de la base d'apprentissage*

3.2. *Algorithmes*

Conformément à ce qui a été expliqué plus tôt, nous avons sélectionné 3 algorithmes de l'état de l'art, que nous appellerons **POP**, **FBC** et **FC**. Dans ce papier, nous proposons également 2 algorithmes hybrides désignés par les sigles **FCDR** et **FCDF**. Le choix et la conception de ces algorithmes ont été motivés par un besoin de personnalisation en temps réel. En effet, il est inenvisageable de faire patienter les participants de notre étude pendant que le système calcule les recommandations. Il a donc fallu choisir des méthodes à la fois rapides et reconnues, et les adapter à notre architecture.

Les informations fournies par les volontaires de l'étude constituent la base de test. Les recommandations sont générées à partir du profil de l'utilisateur courant et de la base d'apprentissage. A aucun moment la base de test n'est intégrée à la base d'apprentissage. De cette manière, les recommandations sont calculées dans les mêmes conditions pour les 250 répondants de l'étude.

POP. Cet algorithme se contente de retourner des items choisis aléatoirement parmi les plus populaires, *i.e.* ceux ayant les votes moyens les plus élevés et un grand nombre de votants. Il s'agit de l'algorithme de référence dans le cadre de notre étude.

FBC. Nous reprenons ici l'un des algorithmes de filtrage basé sur le contenu proposés par (Bradley *et al.*, 2001). Nous sélectionnons les items à recommander en

fonction de leur similarité avec les items appréciés par l'utilisateur courant. Dans ce cas de figure, nous privilégions volontairement la similarité, plutôt que la diversité, pour vérifier si les utilisateurs perçoivent une différence par rapport aux autres algorithmes.

Afin de calculer la similarité entre deux films (équation (1)), nous avons choisi des coefficients de pondération et des mesures de similarité par attribut en fonction de tests d'erreur réalisés sur notre base d'apprentissage. Les coefficients de pondération sur les attributs sont les suivants : $w_{annee} = 0,5$; $w_{realisateur} = 1$; $w_{acteur} = 1$; $w_{genre} = 1,5$; $w_{langue} = 0,25$; $w_{popularite} = 0,5$; $w_{saga} = 1$; $w_{scenariste} = 0,25$. Ainsi, à titre d'exemple, le fait que deux films aient le même réalisateur aura deux fois plus de poids dans le calcul de similarité qu'une année de production équivalente.

Les mesures de similarité par attribut sont définies comme suit : $sim_{acteur}(i_1, i_2) = \frac{\cap_{acteurs}}{\cup_{acteurs}}$; $sim_{genre}(i_1, i_2) = \frac{\cap_{genres}}{\cup_{genres}}$. La similarité pour l'année est égale à 1 si l'écart entre les années de production est inférieur à 5 ans, 0 sinon. La similarité de popularité est égale à 1 s'ils appartiennent à la même classe de popularité (***i.e.*** écart des moyennes des votes inférieur à un seuil fixé et nombres de votants comparables). Enfin, les similarités pour le réalisateur, la langue, la saga et le scénariste valent 1 si les deux films ont la même valeur d'attribut, 0 sinon.

FC. Afin de réduire au maximum le temps de calcul des recommandations en ligne, nous avons opté pour un algorithme de filtrage collaboratif basé sur les items permettant d'effectuer autant de calculs que possibles hors ligne (Sarwar ***et al.,*** 2001). Dans un premier temps, cet algorithme transforme la matrice de votes utilisateur-item en matrice de similarités item-item. Par la suite, il applique une formule pour prédire la note sur un item i non encore voté par l'utilisateur courant, comme étant la moyenne des votes déjà connus pondérée par les similarités entre l'item i et chacun des items contenus dans le modèle de préférences de l'utilisateur. Notre implémentation repose sur la métrique de corrélation de Pearson (Castagnos, 2008). A chaque itération, nous sélectionnons les 10 items avec les notes prédites les plus élevées.

Enfin nous avons imaginé les algorithmes FCDR et FCDF, variantes de l'algorithme FC avec une hybridation par contenu, avec l'objectif de faire fluctuer le degré de diversité des recommandations proposées. Grâce à ces deux alternatives, nous pourrons étudier les éventuels écarts de perception des utilisateurs face à des niveaux de diversité différents.

FCDR. Il s'agit d'un algorithme de sélection hybride, acronyme de Filtrage Collaboratif avec Diversité Relative. Dans un premier temps, nous appliquons l'algorithme FC pour calculer le top-50. Nous plaçons le premier élément du top-50 dans la liste des recommandations. Puis nous y ajoutons les items un par un, en sélectionnant à chaque fois l'item du top-50 qui maximise la fonction de diversité relative par rapport aux recommandations déjà retenues (équation (2)), et ce jusqu'à obtenir le nombre de recommandations souhaité.

FCDF. Cet algorithme, acronyme de Filtrage Collaboratif avec Diversité Fixe, est sensiblement similaire à FCDR, à la différence près qu'un pourcentage fixé (x%) des

recommandations doit provenir uniquement de l'algorithme FC. En d'autres termes, au lieu d'initialiser la liste des recommandations avec le premier élément du top-50, nous retenons les n premiers items du top-50 (avec n = nombre de recommandations souhaité * x%). Dans notre implémentation, ce seuil x a été fixé à 60%.

3.3. *Procédure*

Notre étude est prévue pour durer entre 15 et 20 minutes pour chaque participant. Après une courte page d'accueil pour présenter le contexte de notre étude, chaque volontaire est invité à compléter une procédure en 4 étapes, décrite ci-dessous.

Étape 1. Un premier questionnaire permet de collecter les données démographiques (prénom, nom, email, sexe, âge, nationalité et catégorie socio-professionnelle) et les habitudes cinématographiques de l'utilisateur (fréquence des visites au cinéma, genres cinématographiques appréciés, avec qui il va au cinéma et comment il choisit le film à voir, s'il a pour habitude de lire des sites web, magazines ou romans en lien avec le cinéma). Ces habitudes nous renseignent sur le degré d'expertise de l'utilisateur dans le domaine du cinéma. Nous vérifions également s'il sait ce qu'est un système de recommandation et s'il en a déjà utilisé. Ces questions ont uniquement pour but d'effectuer des statistiques sur les participants de l'étude et d'écarter éventuellement les utilisateurs dont les réponses ne seraient pas pertinentes.

A l'issue de l'étape 1, l'utilisateur est enregistré dans la base de données et le système lui affecte aléatoirement l'un des 5 algorithmes de recommandation disponibles. Les participants de l'étude sont donc implicitement répartis dans 5 groupes, mais cela se fait de manière transparente de sorte qu'ils n'en ont pas conscience. L'affectation se fait de manière à ce qu'il y ait au final le même nombre d'utilisateurs dans chaque groupe.

Étape 2. Lors de cette deuxième phase, le système demande à l'utilisateur de voter pour une série de 100 films sur une échelle de 1 (je déteste) à 5 (j'adore), en prétextant avoir besoin de compléter son modèle de préférences. Ces 100 films lui sont présentés sous forme de 10 pages de 10 films, afin qu'il ne soit pas découragé. Par défaut et dans un souci de synthèse, le système affiche pour chaque film des informations minimales telles que le titre, l'affiche du film, le réalisateur, les genres, les acteurs principaux et l'année. Le participant peut néanmoins obtenir davantage de détails en cliquant sur un lien.

Ce que l'utilisateur ne sait pas, c'est que seuls les 3 premières pages de votes (30 films) sont communes à tous les participants afin d'initialiser son profil. Pour les pages 4 à 10, l'algorithme de recommandation qui a été affecté à l'utilisateur prend le relais et génère à chaque page une liste de 10 films susceptibles d'intéresser cette personne en fonction des informations déjà connues. Sur chaque page de votes, les films sont positionnés dans un ordre aléatoire pour ne pas introduire de biais. Bien sûr, étant donné la taille de la base d'apprentissage (509 films), le risque existe que la qualité des recommandations finisse par décroitre par manque d'items intéressants et

non encore votés. Néanmoins, le risque est assez faible puisqu'ils ne voteront que sur un cinquième de la base. De plus, ce phénomène impacte tous les algorithmes de la même manière dans les différents groupes.

Étape 3. Lors de cette étape, le système propose à l'utilisateur un programme TV d'une semaine (un film par soir) constitué de 5 chaînes. Chaque chaîne correspond en réalité à l'application d'un algorithme (une chaîne représentative de **FC**, une correspondant à **FBC**, etc.). Ici, nous annonçons explicitement à l'utilisateur qu'il s'agit de recommandations (sans expliciter le fonctionnement des différents algorithmes) afin de voir vers quel algorithme se tourne leur confiance, s'ils sont capables de distinguer les listes et si le manque de diversité peut leur poser problème sur une semaine. Pour cela, il doit classer les chaînes par ordre de préférence.

Étape 4. Un post-questionnaire permet à l'utilisateur d'expliciter et de quantifier les performances des algorithmes de l'étape 3. En particulier, nous demandons à l'utilisateur d'évaluer sur une échelle de Likert à 7 modalités les éléments suivants : la pertinence des recommandations, la diversité des films proposés lors des étapes 2 et 3, et son degré de confiance dans le classement des listes qu'il a effectué à l'étape 3.

3.4. *Hypothèses*

Avant de mener à bien cette étude, nous avons listé les hypothèses suivantes :

H1. Les utilisateurs perçoivent la diversité. Les résultats du post-questionnaire devraient donc réfléter cette tendance, en particulier pour les groupes affectés à FCDR et FCDF.

H2. La diversité améliore la satisfaction des utilisateurs. Les votes collectés lors de l'étape 2 devraient donc être plus élevés pour les groupes affectés à FCDR et FCDF.

H3. Les algorithmes basés sur le contenu augmentent la confiance des utilisateurs, contrairement à ceux basés sur la diversité. Les recommandations provenant de l'algorithme FBC devraient donc rencontrer plus de succès lors de l'étape 3.

3.5. *Participants*

250 volontaires, recrutés par le biais des réseaux sociaux et répartis en groupes de 50 personnes (numérotés de G1 à G5), ont réalisé notre étude qui s'est déroulée dans un intervalle de temps d'une semaine. Cet échantillon de population était composé de 114 femmes et 136 hommes, 205 français et 45 étrangers. Tous sauf un ont déclaré aller au cinéma au moins occasionnellement, mais tous ont un intérêt pour les films.

Pour motiver les participants à renseigner correctement les votes sur les films, nous les avons informés que 20 d'entre eux seraient sélectionnés par tirage au sort et recevraient un DVD en accord avec leurs préférences exprimées dans cette étude.

4. Résultats

Afin de valider nos hypothèses, nous avons commencé par analyser les résultats du post-questionnaire (étape 4). En premier lieu, nous avons transformé les réponses en valeurs numériques (de “Pas du tout d’accord = 1” jusqu’à “Tout à fait d’accord = 7”). Par la suite, nous avons calculé les moyennes des réponses pour chacun des groupes G1 à G5 (cf. tableau 2).

N° du groupe (algo. étape 2)	Étape 3 (tous algorithmes confondus)			Moyenne des votes
	Diversité	Pertinence	Confiance	
G1 (POP)	4,64	3,94	4,98	3,49
G2 (FBC)	4,44	3,26	5,34	3,55
G3 (FC)	**5**	4,04	5,32	3,79
G4 (FCDR)	**4,96**	4,1	**5,38**	3,61
G5 (FCDF)	4,88	**4,45**	5,30	3,60

Tableau 2. *Résultats du post-questionnaire en fonction des groupes constitués et moyenne des votes à l’étape 2*

Validation de H1. Les groupes G3 à G5, ayant utilisé les algorithmes à base de filtrage collaboratif pendant l’étape 2 (FC, FCDR et FCDF), ont trouvé les recommandations des 5 algorithmes plus diverses à l’étape 3 (colonne “Diversité” du tableau 2) que les autres groupes. Nous avons utilisé un test de Student pour confirmer la signification statistique de ce résultat (p-value de 0,05 entre G2 et G3, et de 0,07 entre G2 et G4). En outre, seuls 36 utilisateurs du groupe G2 (FBC) ont trouvé la liste des films à voter en étape 2 diverse, contre 45 à 47 utilisateurs sur 50 pour les autres groupes. Les utilisateurs sont donc capables de percevoir la diversité dans les recommandations, même dans les cas où ces recommandations se font de manière implicite (étape 2), ce qui valide notre hypothèse H1.

Validation de H2. Comme prévu dans H2, les votes moyens de l’étape 2 (colonne de droite du tableau 2) sont plus élevés pour les algorithmes de filtrage collaboratif (FC) et de filtrage par diversité (FCDR et FCDF), par rapport à FBC et POP. Cela semble donc confirmer que des algorithmes plus divers améliorent la satisfaction des utilisateurs lorsque les recommandations sont faites de façon implicite. En revanche, la différence de satisfaction reste marginale entre les 3 algorithmes à base de filtrage collaboratif, et plus particulièrement entre FCDR et FCDF. Il semble donc que le degré de diversité importe peu, du moment qu’un seuil minimal est atteint. Ce dernier point restera à éclaircir dans le cadre d’une prochaine étude où nous ferons varier le degré de diversité plus finement et sur un plus grand nombre de recommandations.

Validation de H3. Si la diversité semble améliorer la satisfaction lors de la phase de recommandation implicite (étape 2), les résultats sont beaucoup plus contrastés pendant l’étape 3 où les utilisateurs ont été prévenus du fait qu’il s’agissait de recommandations. Ainsi, nous avons compté dans le tableau 3 le nombre de fois que chaque algorithme a été positionné en premier choix lors de l’étape 3 (programme TV préféré de l’utilisateur). Tous groupes confondus, nous constatons que l’algorithme FBC rencontre le plus de suffrages. Cela confirme notre intuition de l’hypothèse H3, selon

laquelle le filtrage par contenu suscite davantage de confiance chez l'utilisateur (cf. dernière ligne du tableau 3). Les commentaires laissés par les volontaires en fin d'étude fournissent un début d'explication : grâce aux similarités par attribut, d'aucun a beaucoup plus de facilité à appréhender le lien entre les préférences exprimées et les recommandations fournies par FBC, en comparaison avec les autres algorithmes. Chaque utilisateur peut donc facilement imaginer une explication implicite pour une recommandation (ex. : j'ai mis un vote élevé pour Matrix 1, ce qui explique qu'on me recommande Matrix 2).

N° du groupe	Algorithme choisi à l'étape 3				
	POP	FBC	FC	FCDR	FCDF
G1 (POP)	14	**22**	7	3	4
G2 (FBC)	9	**29**	7	5	**0**
G3 (FC)	7	**17**	16	6	4
G4 (FCDR)	9	**15**	6	12	7
G5 (FCDF)	**14**	10	8	5	12
Confiance (tous util. confondus)	4,98	**5,34**	5,32	**5,34**	5,32

Tableau 3. *Répartition (Nb de personnes) du programme TV préféré à l'étape 3*

En revanche, d'après la colonne "Pertinence" du tableau 2, les groupes G4 et G5 affectés à nos algorithmes de diversité hybrides ont trouvé les recommandations de l'étape 3 plus pertinentes (tous algorithmes confondus) avec plus d'un point de différence par rapport au groupe G2 affecté à l'algorithme par contenu. Ce résultat est statistiquement significatif avec une confiance de 99% ($p = 0,004$ entre G2 et G4 et $p = 4,27e-05$ entre G2 et G5). Fournir des items divers pendant l'étape 2 (FCDR) a également amélioré le degré de confiance global des utilisateurs dans les recommandations, même si le choix de l'algorithme à l'étape 3 se porte majoritairement sur FBC. Par conséquent, quel que soit l'algorithme de recommandation utilisé, il faut s'assurer que le modèle de préférences de l'utilisateur contient des items suffisamment divers pour obtenir de meilleures recommandations. Il s'agit là d'un effet inattendu de la diversité qui peut déboucher sur des travaux en lien avec le choix des items à faire voter pendant la phase de démarrage à froid.

5. Conclusion et Perspectives

Ce travail constitue une étude exploratoire du rôle et de l'impact de la diversité dans les systèmes de recommandations. Il a permis d'illustrer la nécessité de construire des modèles de préférences contenant des items variés pour assurer un bon niveau de pertinence et de confiance dans les recommandations. Par ailleurs, nous avons prouvé que la diversité est perçue et améliore la satisfaction des utilisateurs. Néanmoins, la diversité dans les recommendations peut nécessiter davantage d'explications auprès de l'utilisateur qui ne comprend pas toujours le lien avec ses préférences exprimés. En résumé, la diversité est une dimension complexe qui s'avère bénéfique pour l'utilisateur si elle est utilisée au bon moment et de la bonne manière. Suite à ces

conclusions, une perspective consistera à étudier les moyens de garantir la diversité lors de la phase de démarrage à froid.

6. Bibliographie

Adomavicius G., Kwon Y., « Overcoming Accuracy-Diversity Tradeoff in Recommender Systems: A Variance-Based Approach », *In Proceedings of the 18th Workshop on Information Technology and Systems (WITS'08)*, Paris, France, 2008.

Adomavicius G., Kwon Y., « Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques », *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, n° 5, p. 896-911, 2012.

Agrawal R., Gollapudi S., Halverson A., Ieong S., « Diversifying search results », *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM'09, ACM, Barcelona, Spain, p. 5-14, 2009.

Barragáns-Martínez A. B., Costa-Montenegro E., Burguillo J. C., Rey-López M., Mikic-Fonte F. A., Peleteiro A., « A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition », *Information Sciences*, vol. 180, p. 4290-4311, 2010.

Bradley K., Smyth B., « Improving Recommendation Diversity », *Irish Conference on Artificial Intelligence and Cognitive Science (AICS'01)*, Dublin, Ireland, p. 85-94, 2001.

Castagnos S., Modélisation de comportements et apprentissage stochastique non supervisé de stratégies d'interactions sociales au sein de systèmes temps réel de recherche et d'accès à l'information, Thèse de doctorat, LORIA - Université Nancy 2, 2008.

Castagnos S., Jones N., Pu P., « Recommenders' Influence on Buyers' Decision Process », *In proc. of the 3rd ACM Conference on Recommender Systems (RecSys'09)*, New York, USA, p. 361-364, October, 2009.

Castagnos S., Jones N., Pu P., « Eye-Tracking Product Recommenders' Usage », *Proceedings of the 4th ACM Conference on Recommender Systems (RecSys 2010)*, Barcelona, Spain, September, 2010.

Clarke C. L., Kolla M., Cormack G. V., Vechtomova O., Ashkan A., Buttcher S., MacKinnon I., « Novelty and diversity in information retrieval evaluation », *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR'08, ACM, p. 659-666, 2008.

Ge M., Gedikli F., Jannach D., « Placing High-Diversity Items in Top-N Recommendation Lists », *Workshop on Intelligent Techniques for Web Personalization and Recommender Systems (ITWP'11)*, Barcelona, Spain, 2011.

Häubl G., Murray K., « Preference Construction and Persistence in Digital Marketplaces: The Role of Electronic Recommendation Agents », *Journal of Consumer Psychology*, vol. 13, n° 1, p. 75-91, 2003.

Herlocker J. L., Konstan J. A., Terveen L. G., John, Riedl T., « Evaluating collaborative filtering recommender systems », *ACM Transactions on Information Systems*, vol. 22, p. 5-53, 2004.

Hu R., Pu P., « Helping Users Perceive Recommendation Diversity », *In Proceedings of the 1st International Workshop on Novelty and Diversity in Recommender Systems (DiveRS'11)*, Chicago, USA, 2011.

Jones N., User Perceived Qualities and Acceptance of Recommender Systems, Phd thesis, Ecole Polytechnique Fédérale De Lausanne, 2010.

Koren Y., Bell R. M., Volinsky C., « Matrix Factorization Techniques for Recommender Systems », *IEEE Computer*, vol. 42, n° 8, p. 30-37, 2009.

Lathia N. K., Evaluating Collaborative Filtering Over Time, Phd thesis, University College London, 2010.

McGinty L., Smyth B., « On the Role of Diversity in Conversational Recommender Systems », *International Conference on Case-Based Reasoning (ICCBR'03)*, p. 276-290, 2003.

McNee S. M., Riedl J., Konstan J. A., « Being accurate is not enough: how accuracy metrics have hurt recommender systems », *CHI '06: CHI '06 extended abstracts on Human factors in computing systems*, ACM, Montréal, Canada, p. 1097-1101, 2006.

McSherry D., « Diversity-conscious retrieval », *Proceedings of the 6th European Conference on Advances in Case-Based Reasoning (ECCBR'02)*, London, UK, p. 219-233, 2002.

Radlinski F., Bennett P. N., Carterette B., Joachims T., « Redundancy, diversity and interdependent document relevance », *SIGIR Forum*, vol. 43, n° 2, p. 46-52, 2009.

Ricci F., Rokach L., Shapira B., Kantor P. B., *Recommender Systems Handbook*, Springer, 2011.

Said A., Kille B., Jain B. J., Albayrak S., « Increasing Diversity Through Furthest Neighbor-Based Recommendation », *Proceedings of the WSDM'12 Workshop on Diversity in Document Retrieval*, Seattle, USA, 2012.

Sarwar B., Karypis G., Konstan J., Reidl J., « Item-based collaborative filtering recommendation algorithms », *World Wide Web*, p. 285-295, 2001.

Sawers P., « Remember Netflix's $1m algorithm contest? Well, here's why it didn't use the winning entry », http://thenextweb.com/media/2012/04/13/remember-netflixs-1m-algorithm-contest-well-heres-why-it-didnt-use-the-winning-entry/, 2012.

Schickel V., Faltings B., « Using an Ontologcial A-priori Score to Infer User's Preferences », *Workshop on Recommender Systems, in Conjunction with the 17th European Conference on Artificial Intelligence (ECAI 2006)*, Riva del Garda, Italy, August, 2006.

Sill J., Takacs G., Mackey L., Lin D., Feature-Weighted Linear Stacking, Netflix prize report, Cornell University, 2009.

Smyth B., McClave P., « Similarity vs. Diversity », *Proceedings of the 4th International Conference on Case-Based Reasoning*, Vancouver, Canada, p. 347-361, 2001.

Töscher A., Jahrer M., The BigChaos Solution to the Netflix Grand Prize, Netflix prize report, Commendo Research, 2009.

Wan S., Xue Y., Yu X., Guan F., Liu Y., Cheng X., « ICTNET at Web Track 2011 Diversity Track », *Text REtrieval Conference (TREC'11)*, 2011.

Yu C., Lakshmanan L. V., Amer-Yahia S., « Recommendation Diversification Using Explanations », *Proceedings of the 2009 IEEE International Conference on Data Engineering (ICDE'09)*, p. 1299-1302, 2009.

Zhang M., Hurley N., « Avoiding Monotony: Improving the Diversity of Recommendation Lists », *Proceedings of the 2nd ACM Recommender Systems*, Lausanne, Switzerland, p. 123-130, 2008.

Ziegler C.-N., McNee S., Konstan J., Lausen G., « Improving recommendation lists through topic diversification », *Proceedings of the 14th international conference on World Wide Web (WWW'05)*, p. 22-32, 2005.