Utilisation d'invariants pour une médiation inter-domaines de modèles utilisateurs : ressources invariantes et invariants sémantiques

Emilien Perrin, Armelle Brun, Anne Boyer

Loria - Nancy-Université - Équipe Kiwi
Campus Scientifique - BP 239
54506 Vandoeuvre-lès-Nancy Cedex, France
{emilien.perrin, armelle.brun, anne.boyer}@loria.fr
http://kiwi.loria.fr

Résumé. Les services de personnalisation du Web 2.0 reposent sur l'exploitation de modèles utilisateurs. Schématiquement, plus la quantité d'informations sur les utilisateurs est grande, meilleures sont la modélisation et la qualité du service. En pratique, nombre de services rencontrent un problème de manque d'informations sur les utilisateurs. Dans cet article, nous y répondons par médiation inter-domaines de modèles utilisateurs, c'est-à-dire la complétion de modèles en exploitant des données d'un autre domaine. La médiation que nous proposons repose sur un transfert d'informations inter-domaines. Ce transfert consiste en l'utilisation de couples invariants ou très corrélés pouvant être des couples de ressources ou de descripteurs sémantiques, identifiés après enrichissement sémantique des modèles. Nous montrons que le transfert sous forme de couple de ressources permet une complétion de qualité et que l'exploitation de descripteurs sémantiques augmente la couverture à qualité égale. Enrichir sémantiquement est donc bénéfique pour le transfert inter-domaines.

1 Introduction

De plus en plus d'informations sont mises à disposition des utilisateurs sur le Web et pour éviter qu'ils ne soient perdus, des services de personnalisation leurs sont proposés : personnalisation du contenu, de l'agencement, recommandation de ressources, etc. (Adomavicius et al., 2008). Pour fournir un service personnalisé à un utilisateur, dit utilisateur courant, le système de personnalisation doit connaître cet utilisateur. Il repose pour cela sur un modèle de l'utilisateur. Cette modélisation peut porter sur divers aspects de l'individu en fonction de la personnalisation fournie : ses compétences, ses préférences, etc. Un modèle peut prendre plusieurs formes : des variables binaires ou pondérées jusqu'à des représentations structurées de connaissances. En règle générale, plus le système de personnalisation considéré possède d'informations sur l'utilisateur courant, meilleure est la qualité du service fourni (dans la limite d'un certain seuil (Lam et Riedl, 2005)). Dans la pratique cependant, la modélisation de l'utilisateur peut être très incomplète, voire inexistante, notamment en cas de nouvelle interaction

avec le système. Un des nombreux défis dans la littérature est donc d'augmenter la quantité d'informations disponible sur les utilisateurs de façon à améliorer la qualité de la modélisation dans le système, que nous appellerons système cible.

Avec la démocratisation du Web et notamment 2.0, un nombre croissant de systèmes modélise ses utilisateurs. Dans un but de compléter la modélisation de l'utilisateur courant dans le système cible, la solution envisagée dans cet article est d'exploiter des données de modélisation sur l'utilisateur courant issues d'autres systèmes, dits systèmes sources. Cette exploitation est appelée médiation de modèles utilisateurs (Berkovsky et al., 2007). Nous proposons d'effectuer une médiation inter-domaines où il pourra s'agir, par exemple, d'utiliser des données de modélisation d'un utilisateur dans le domaine des publicités pour aider à la modélisation de cet utilisateur dans le domaine de la musique. Notons que cette approche nécessite d'être vigilant quant à la protection de la vie privée des utilisateurs.

La médiation inter-domaines de modèles utilisateurs que l'on peut définir par l'utilisation de données de modélisation issues d'un autre domaine, est un axe de recherche récent et en plein essor depuis les travaux de Berkovsky en 2006 (Berkovsky, 2006), en particulier nous pouvons citer Li et al. (2009); Zhang et al. (2010); Pan et al. (2010). Cependant, ces modèles ont l'inconvénient d'être difficilement compréhensibles par un humain car ils ne cherchent pas à identifier les informations directement transposables entre les domaines.

Partant du constat que les travaux les plus récents reposent sur l'exploitation d'invariants entre domaines, nous proposons une approche de médiation construisant un modèle compréhensible par un humain et qui exploite les informations directement transposables. Dans ce cadre, nous choisissons d'identifier et d'exploiter des couples de ressources entre domaines (une ressource dans le domaine source, une ressource dans le domaine cible), au sein desquels les appréciations des utilisateurs sont invariantes ou fortement corrélées. Ces couples seront exploités dans le but de compléter les modèles utilisateurs dans le domaine cible lorsque des informations sont connues dans le domaine source. Nous poussons cette idée un peu plus loin en utilisant des informations sémantiques caractérisant les ressources. La sémantique n'a, à notre connaissance, été que très peu exploitée pour la médiation inter-domaines. L'approche présentée dans Shi et al. (2011) est la seule exploitant la sémantique : elle utilise les tags attribués aux ressources par les utilisateurs. Nous sommes convaincus que l'exploitation de la sémantique permettra de mieux transférer des informations entre des domaines très éloignés conceptuellement ou relevant de champs d'activités humaines éloignés, que la simple utilisation de couple de ressources corrélées. Par exemple, une technique simple établira que les personnes aimant telle publicité R_i dans le domaine de la publicité (domaine source), aiment également telle musique R_j dans le domaine de la musique (domaine cible). Introduire des informations sémantiques sur les ressources permettra de trouver qu'un lien entre R_i et R_j est le compositeur de la musique R_j utilisée dans la publicité R_i. Ainsi, lorsque le système saura qu'un utilisateur apprécie une publicité dont la bande son a été composée par ce compositeur, alors il pourra compléter le modèle de cet utilisateur dans le domaine de la musique, en renseignant qu'il apprécie les musiques de ce compositeur. De cette manière, le transfert d'information entre domaines *a priori* éloignés conceptuellement pourra être possible. L'exploitation de la sémantique a également un avantage certain pour l'expert qui voudra interpréter la connaissance transférée entre domaines et qui manipulera des variables explicatives plus facilement interprétables.

Un état de l'art de la médiation inter-domaines de modèles utilisateurs est présenté en

section 2. La section 3 s'intéresse au modèle de médiation inter-domaines que nous proposons. La section 4 présente les expériences que nous avons menées pour valider notre approche. Enfin, la section 5 conclut ce travail et présente quelques perspectives.

2 État de l'art

La personnalisation, domaine de recherche en plein essor, est "la capacité à fournir des contenus et services adaptés aux utilisateurs en fonction de la connaissance dont l'on dispose à propos de leurs préférences et de leur comportement" [1]. Cette personnalisation peut se décliner en publicité ciblée, filtrage de l'information, recommandation, personnalisation de l'interface, etc. Elle consiste en le processus itératif suivant (Adomavicius et Tuzhilin, 2005) : compréhension des besoins utilisateur, livraison du service personnalisé exploitant une modélisation utilisateur, mesure de satisfaction de l'utilisateur.

Les données stockées dans un modèle utilisateur doivent être pertinentes eu égard au service fourni. Elles peuvent donc porter sur tout ou partie des catégories suivantes : les identifiants, les préférences et intérêts généraux, les compétences, les caractéristiques individuelles (relatives à l'âge ou au handicap en particulier), les intentions/but(s) courant(s), les croyances de l'individu, son état psychologique, ainsi que les interactions avec le système et les régularités comportementales de l'utilisateur qui peuvent en être déduites (Jameson, 1999). De plus le contexte d'interaction peut être pris en compte (Berkovsky et al., 2007) (utilisation le soir, avec des amis, sur un téléphone mobile, etc.). Les formats des modèles sont très souvent propriétaires et leur représentation est très variable : bases de données, fichier plat, etc. et différents formalismes de représentation peuvent être utilisés. Kostadinov (2006) identifie notamment : les profils à base de mots clés ou prédicats pondérés, les profils à base de formules, les profils multidimensionnels et à base d'ontologies.

Les données qui composent un modèle utilisateur peuvent être recueillies explicitement (ce sont les saisies, l'évaluation de ressources faites par les utilisateurs, etc.) ou implicitement (à l'aide de tests ou monitorage de l'usage que les utilisateurs font du service). Un problème rencontré par de nombreux systèmes est le manque de données sur l'utilisateur (Zhou et Luo, 2009), ce qui limite la qualité de la modélisation et donc du service. En effet, en cas de nouvelle interaction avec un système, les données sur l'utilisateur sont peu importantes. Cependant, demander à l'utilisateur de saisir des informations le concernant (via un questionnaire par exemple) n'est pas une solution puisque ce dernier aura probablement déjà saisi des informations le concernant dans d'autres systèmes et ne souhaitera probablement pas recommencer.

Certains travaux se sont intéressés à l'analyse des préférences utilisateurs et notamment aux facteurs permettant de comprendre et de déduire des préférences. Dans ce cadre, il a été prouvé que des invariants sous-tendant les préférences des utilisateurs au sein d'un domaine peuvent être utilisés pour caractériser les préférences des utilisateurs (Hofmann et Puzicha, 1999). Par ailleurs, certains travaux supposent que des caractéristiques propres à un utilisateur sont invariantes entre domaines (Li et al., 2009; Pan et al., 2010). Si l'on considère que chaque système de personnalisation se rapporte à un domaine, et en se basant sur la supposition précédente, il est donc pertinent d'exploiter l'information issue d'un système où un utilisateur est

1. Paul Hagen, Forrester Research, 1999.

déjà modélisé de façon à améliorer sa modélisation dans le système cible et ainsi pallier le manque de données.

En 2006 et 2007, Berkovsky et al. (2007) ont introduit le terme de médiation dans le cadre de la personnalisation. La médiation est l'import et l'intégration par un système, de données collectées par d'autres systèmes. Ils ont formalisé un cadre de médiation et défini plusieurs modes de médiation inter-domaines. La médiation se fait au travers d'un médiateur, entre un système (dit cible), qui manque d'informations et d'autres systèmes (dits sources) qui possèdent des données de modélisation sur l'utilisateur. Ce dernier est composé d'un mécanisme d'intégration (en charge de la résolution de conflits et de l'hétérogénéité) et d'une base de connaissance. Dans le cadre de systèmes de recommandation, les auteurs distinguent :

- l'*approche standard* où les systèmes sources envoient l'ensemble de leurs modèles utilisateurs locaux (dans ce cas, des vecteurs de votes) au système cible, suite à quoi le système cible réalise une matrice unifiée. Cette méthode pose dans les faits de sérieux problèmes relativement à la protection de la vie privée des individus et ne pourrait être mise en œuvre que dans une fédération de systèmes pour lesquels l'échange de données personnelles serait autorisé. De plus, elle est coûteuse en terme d'échange de données ;
- l'*approche inter-domaines* dans laquelle, suite à une requête du système cible, les systèmes sources envoient au système cible la liste des utilisateurs similaires ainsi que leurs valeurs de similarité ;
- l'*approche de distance à la moyenne*, propre au cadre de la recommandation, dans laquelle les systèmes sources envoient leurs recommandations au système cible qui les intègre en une seule liste de recommandations.

Suite à ces travaux, plusieurs techniques sur la médiation inter-domaines ont été proposées et peuvent être classées selon l'approche utilisée :

- *Le transfert de codebook* (Li et al., 2009). Un regroupement (bi-cluster) utilisateur-ressource homogène (en fonction des notes dans la matrice d'évaluations $Utilisateur \times Ressource \rightarrow note$) est effectué dans les domaines sources. En identifiant les centres de ces bi-clusters, un codebook [2] est déterminé, permettant de remplir la matrice de notes dans le domaine cible par expansion du codebook. Les valeurs manquantes sont remplacées en fonction de leur proximité attendue avec les valeurs connues. Le transfert se fait en dupliquant certaines lignes et colonnes du codebook, en supposant qu'un même comportement peut être retrouvé dans les deux domaines. Cette technique a, contrairement aux précédentes techniques, l'avantage de ne pas nécessiter que les utilisateurs ou les ressources soient les mêmes dans les deux domaines.
- *La décomposition de matrices*. Ces techniques transforment à la fois les ressources et les utilisateurs dans le même espace latent de description, les rendant directement comparables. L'espace de description latent essaie d'expliquer les votes en caractérisant à la fois les ressources et les utilisateurs Koren (2008). On trouve les approches de Zhang et al. (2010) où les utilisateurs ne sont pas nécessairement partagés et de Pan et al. (2010) où les utilisateurs doivent être partagés.

Certaines de ces techniques sont exploitées dans le but de remplir l'ensemble des modèles utilisateurs dans le domaine cible et non à transférer le seul modèle de l'utilisateur courant, ce qui se traduit par un coût important en terme de bande passante.

2. Un codebook est une fonction de traduction exprimée par ses correspondances en entrée et en sortie.

Un constat peut être tiré de l'analyse de ces différentes approches : elles ne cherchent pas à identifier directement les informations (quelle que soit leur nature) invariantes entre les domaines. De plus, aucune n'utilise d'informations sémantiques sur les ressources (qui sont à disposition sur le Web actuel, issues de DBpedia ou IMDb par exemple, ou dans les bases de données ou de connaissances internes). Nous sommes convaincus que le transfert d'informations entre domaines sera facilité par l'identification de concepts sémantiques invariants chez les utilisateurs. Ces deux constats nous ont amené à investiguer la recherche d'informations invariantes entre domaines et la voie de l'enrichissement sémantique des informations contenues dans les modèles utilisateur à l'aide de données externes.

Récemment, une approche exploitant des informations sémantiques a été proposée (Shi et al., 2011), elle utilise les mots-clés (tags) saisis par les utilisateurs communs aux différents domaines. Elle se base sur une factorisation de matrices avec la contrainte qu'il y ait des similarités entre utilisateurs et ressources à travers les domaines. Cette approche repose sur les tags attribués aux ressources par les utilisateurs. C'est en ce sens qu'elle est différente de l'approche que nous proposons et que nous présentons dans la section ci-dessous.

3 Exploitation de couples corrélés pour la médiation inter-domaines de modèles utilisateurs

Soit U l'ensemble des utilisateurs, divisé en deux sous-ensembles : les utilisateurs d'apprentissage $U_{appr} = \{u_{a1}, \cdots, u_{ak}\}$ et les utilisateurs de test $U_{test} = \{u_{t1}, \cdots, u_{tl}\}$. $R_s = \{r_{s1}, \cdots, r_{sm}\}$ est l'ensemble des ressources dans le domaine source et $R_c = \{r_{c1}, \cdots, r_{cp}\}$ est l'ensemble des ressources dans le domaine cible. Les modèles utilisateurs, dans le domaine source ($u_{ai}(r_{s1}, \cdots, r_{sm})$ et $u_{tj}(r_{s1}, \cdots, r_{sm})$) ainsi que dans le cible ($u_{ai}(r_{c1}, \cdots, r_{cp})$ et $u_{tj}(r_{c1}, \cdots, r_{cp})$), sont des modèles d'appréciation sur les ressources dans lesquels les appréciations (par exemple $u_{ai}(r_{s1})$) sont des valeurs numériques. Notons que les utilisateurs (d'apprentissage et de test) sont communs aux deux domaines source et cible. Cette contrainte peut sembler forte mais il s'agit d'une situation trouvable dans le cadre de systèmes interopérants.

L'approche que nous proposons cherche à identifier des ressources corrélées entre domaines, et donc à des appréciations transférables. Dans ce cadre, nous nous intéressons dans un premier temps à l'identification de couples de ressources invariantes ou fortement corrélées.

3.1 Approche par couples de ressources corrélées

3.1.1 L'identification de couples corrélés

L'identification de couples de ressources corrélées entre domaines repose sur l'exploitation de données d'apprentissage : des modèles d'appréciation des utilisateurs d'apprentissage U_{appr} dans le domaine source et dans le domaine cible (les deux parties gris clair dans la figure 1).

Rappelons qu'un couple de ressources est composé d'une ressource du domaine source et d'une ressource du domaine cible. Pour chaque couple possible de ressources, la valeur de la corrélation entre ces deux ressources est évaluée. La mesure de corrélation que nous avons choisie est la corrélation de Pearson, classiquement utilisée dans le domaine de la modélisation utilisateur : cette corrélation représente dans quelle mesure les utilisateurs notent ces deux ressources de façon similaire (Breese et al., 1998). La figure 1 présente les domaines source et

cible ainsi que leurs ressources respectives, la répartition entre utilisateurs d'apprentissage et de test, ainsi que des exemples de couples de ressources dont on évalue la corrélation.

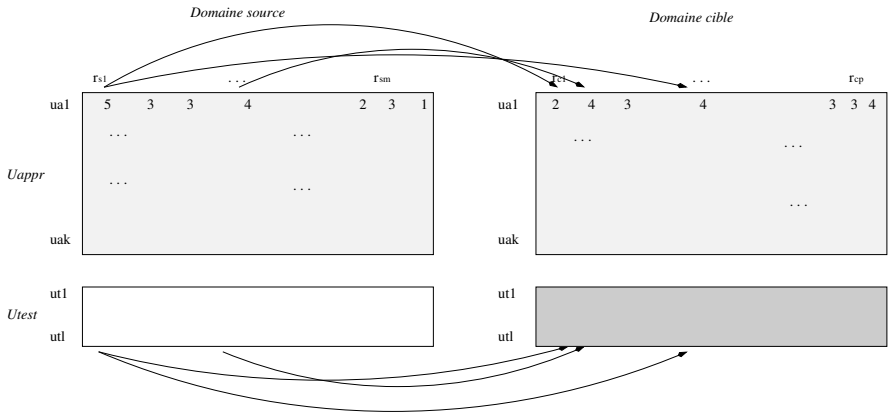

FIG. 1 – *Principe général de la recherche et de l'exploitation de couples corrélés*

Une fois les corrélations apprises, un seuil de corrélation minimale est fixé et l'ensemble C composé de couples de ressources (r_s, r_c) dont la corrélation est supérieure à ce seuil est construit. Cet ensemble servira à compléter les modèles utilisateurs dans le domaine cible.

3.1.2 La complétion de modèles

Sachant l'ensemble C de couples retenus dans l'étape précédente, et les modèles des utilisateurs de test dans le domaine source (partie blanche de la figure 1), nous cherchons à compléter les modèles de ces utilisateurs dans le domaine cible (partie gris foncé de la figure 1).

Pour un utilisateur de test u_{ti}, nous estimons son appréciation pour la ressource r_c du domaine cible ($u_{ti}(r_c)$) comme étant la moyenne des appréciations de l'utilisateur u_{ti} sur les ressources du domaine source fortement corrélées avec r_c et appartenant à l'ensemble C, comme montré en équation 1.

$$u_{ti}(r_c) = \frac{\sum_{r_{sa},(r_{sa},r_c)\in Cr_s} u_{ti}(r_{sa})}{|Cr_c|} \quad (1)$$

où r_{sa} représente une ressource dans le domaine source et Cr_c représente l'ensemble des couples de ressources corrélées dont l'élément du domaine cible est la ressource r_c.

L'évaluation de la qualité de la complétion se fait en calculant l'erreur moyenne entre l'appréciation estimée (équation 1) et l'appréciation effective dans des utilisateurs de test dans le domaine cible : la MAE (Mean Absolute Error).

3.2 Approche par enrichissement sémantique

Nous nous intéressons maintenant à l'exploitation d'informations sémantiques sur les ressources de façon à extraire des caractéristiques sémantiques corrélées entre les domaines. Avec

la même approche que dans la section précédente, nous cherchons à identifier des couples descripteurs sémantiques invariants entre domaines.

Nous effectuons dans un premier temps un enrichissement sémantique (Berkovsky et al., 2007) des modèles d'appréciation sur les ressources (domaine source et domaine cible). Cet enrichissement se fait à l'aide d'une base de données ou de connaissances de chaque domaine. Le résultat se présentera sous la forme de modèles d'appréciation des utilisateurs sur l'ensemble des descripteurs sémantiques des ressources. L'ensemble des descripteurs sémantiques dans un domaine (source ou cible) est constitué de l'union de l'ensemble des descripteurs sémantiques des ressources de ce domaine. L'ensemble des descripteurs du domaine source sera noté $D_s = d_{s1}, \cdots, d_{sn}$ et l'ensemble des descripteurs du domaine cible sera noté $D_c = d_{c1}, \cdots, d_{cq}$. Pour un utilisateur u_{ti} et un descripteur d_{sj} donnés, nous estimons la valeur de l'appréciation $u_{ti}(d_{sj})$ comme étant la moyenne des appréciations de cet utilisateur sur les ressources comportant ce descripteur, (équation 2).

$$u_{ti}(d_{sj}) = \frac{\sum_{r_{sb} \in R_{d_{sj}} \ et \ u_{ti}(r_{sb}) != ?}}{|R_{d_{sj}}|} \tag{2}$$

où $R_{d_{sj}}$ est l'ensemble des ressources qui ont d_{sj} comme descripteur sémantique et r_{sb} itère sur les ressources de cet ensemble et pour lesquelles le système connaît l'appréciation de l'utilisateur u_{ti} ($u_{ti}(r_{sb})! =?$).

La recherche de couples corrélés s'effectue de façon identique à l'approche par ressources, mais ici les couples extraits sont des couples de descripteurs. De la même façon, l'étape de complétion des modèles s'effectue de façon similaire à l'approche par ressources (équation 1).

4 Expériences

Nous présentons dans cette partie les différentes expériences que nous avons menées dans le but de valider notre approche par couples (de ressources ou de descripteurs sémantiques) invariants et fortement corrélés.

4.1 Choix des données d'expérimentation

Le choix des données d'expérimentation a consisté en le choix du corpus, le choix des domaines et le choix des utilisateurs.

A notre connaissance, aucun jeu de données publiques modélisant les préférences des mêmes utilisateurs dans deux domaines différents et pouvant être enrichi avec des informations extérieures n'existe librement. Nous avons donc choisi d'utiliser un jeu de données que nous partitionnons en sous-domaines de façon à simuler deux domaines différents, comme cela a été fait dans Berkovsky et al. (2007). Le jeu de données que nous avons choisi est celui issu du système de recommandation MovieLens : MovieLens10M. Ce corpus est composé de dix millions d'appréciations (sur une échelle de 1 à 5) sur 10.681 films par 71.567 utilisateurs ayant évalué au moins 20 films. Nous avons choisi ce jeu car il peut être enrichi à l'aide des informations de la base de données publique sur les films IMDb (Internet Movie Database). D'autre part, les utilisateurs ont été sélectionnés aléatoirement et les chances de trouver des utilisateurs aux profils différents sont donc maximisées.

Les deux domaines simulés, ici les genres des films, sont choisis de façon à être le plus éloignés possible. La sélection se fait selon leur proximité dans la matrice de corrélation de matrices de factorisation des matrices de votes (Zhang et al., 2010). Les deux genres les plus éloignés sont Thriller et Comédie et nous choisissons d'effectuer le transfert de Thriller vers Comédie. Les genres se recouvrent totalement en terme d'utilisateurs et partiellement en terme de films évalués, comme dans le cas du e-commerce où une ressource peut appartenir simultanément à deux domaines. Le nombre de films considérés dans le domaine Thriller est de 1.672, il est de 3.138 pour le domaine Comédie.

20% des utilisateurs, soit 5.671, ont été choisis aléatoirement parmi les utilisateurs ayant évalué au moins 20 films dans les genres Thriller et Comédie.

4.2 Préparation des données

Pour préparer les données, nous avons réalisé un enrichissement sémantique et un filtrage des descripteurs sémantiques obtenus.

Pour effectuer l'enrichissement sémantique, nous avons réalisé une correspondance sur les titres. Les modèles enrichis ont ensuite été générés en exploitant l'équation (2). Les types d'attributs retenus sont : les mots-clés, le lieu de tournage, le pays producteur, les durées du film, les acteurs, les scénaristes et les compositeurs.

Pour réduire la dimensionnalité et éviter que des descripteurs ne soient évalués par trop peu d'utilisateurs, nous avons filtré les descripteurs. Nous avons retenu les mots-clés caractérisant au moins 50 films, les personnes (acteur, compositeur, etc.) intervenant dans au moins 6 films différents. De plus, si la personne est acteur, elle n'est comptabilisée que si elle est tête d'affiche (c'est-à-dire dans les 5 premiers acteurs du film dans l'IMDb). Le nombre de descripteurs résultant dans le domaine Comédie est de 3.154, il est de 3.138 pour le domaine Thriller.

4.3 Résultats

Pour évaluer les deux méthodes que nous proposons et les conditions d'utilisation en faveur de l'une ou de l'autre, les modèles que nous complétons (domaine cible) sont entièrement vidés et nous avons complété ces modèles utilisateurs en exploitant les couples appris dans les sections 3.1 et 3.2. La qualité de la complétion sera évaluée à la fois en terme de MAE en comparant les appréciations effectivement mises par les utilisateurs et les appréciations estimées, mais également en terme de couverture : quel pourcentage du modèle est complétée.

4.3.1 Couples de ressources corrélées (Rs vers Rc)

La première expérience que nous menons cherche à évaluer la qualité des modèles complétés lorsque des couples de ressources sont exploités pour effectuer le transfert entre les deux domaines (de R_s vers R_c). Les modèles que nous cherchons à remplir sont les modèles des utilisateurs U_{test} du domaine cible Comédie. Ces modèles sont en moyenne remplis avec 3,1% d'appréciations.

Les courbes continues de la figure 2 présentent la MAE et la couverture obtenues en fonction du seuil au-dessus duquel les couples de ressources sont considérés comme corrélés. Nous pouvons remarquer que, comme attendu, plus le seuil de corrélation est élevé, plus la précision de la prédiction dans le domaine cible s'améliore (la MAE diminue). En effet, plus la valeur

du seuil est élevée, plus les couples retenus sont composés de ressources très corrélées et donc plus le transfert est de bonne qualité. Cependant, cela se fait au détriment de la couverture qui diminue. On peut notamment voir que la couverture est inférieure à 1% lorsque le seuil est de 0,9. On peut aussi noter qu'un taux de complétion équivalent au taux d'origine (2,8% contre 3,1% à l'origine) est obtenu pour un seuil inférieur ou égal à 0,7. D'autre part, dans la mesure où la moyenne des appréciations est similaire dans les deux domaines considérés (3,51 pour comédie et 3,58 pour thriller), nous avons cherché à savoir si notre méthode se comportait mieux que si l'on complétait les modèles utilisateur dans le domaine cible par les moyennes individuelles d'appréciation de l'ensemble des ressources (films) dans le domaine source. La MAE ainsi obtenue plus de deux fois plus élevée que celle de notre méthode (0,783 contre 0,361 pour un seuil de corrélations à 0,7 par exemple, l'écart se resserrant lorsque le seuil diminue). Nous pouvons conclure que notre méthode est plus précise qu'un simple transfert de l'appréciation moyenne des utilisateurs.

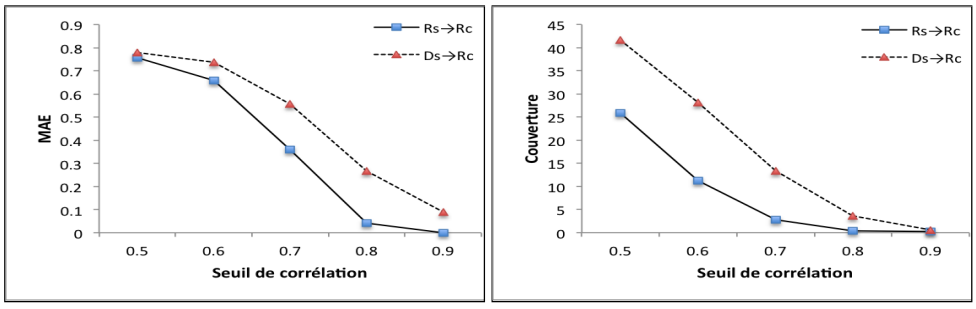

FIG. 2 – *MAE et couverture de la complétion du domaine cible non enrichi*

4.3.2 Enrichissement du domaine source (Ds vers Rc)

Nous nous intéressons à présent à la complétion de modèles des utilisateurs U_{test} dans le domaine cible lorsque les modèles du domaine source sont enrichis sémantiquement et que les modèles du domaine cible ne le sont pas. L'évolution de la MAE et de la couverture en fonction du seuil de corrélation est présentée dans la figure 2 (courbes en pointillés).

Nous pouvons remarquer que comme précédemment, plus le seuil de corrélation est élevé, plus la MAE dans le domaine cible est faible. D'autre part, une couverture équivalente au taux de complétion des modèles utilisateurs originaux dans le domaine cible (3,6% contre 3,1% originellement) est atteinte dès que le seuil est inférieur à 0,8, contre 0,7 pour le passage de R_s vers R_c. La MAE obtenue avec enrichissement sémantique du domaine source est toujours supérieure à celle obtenue par utilisation des modèles utilisateurs non enrichis (qualité moindre). Cependant, cette perte est associée à une couverture bien plus élevée que sans enrichissement (amélioration de 15 points pour un seuil de 0,5). On notera que pour un seuil de corrélation élevé, les couvertures sont comparables et extrêmement basses.

Comme précédemment, nous avons complété les modèles de chaque utilisateur avec leur moyenne d'appréciation dans le domaine source. La MAE ainsi obtenue est plus élevée qu'avec

notre méthode (0,797 pour l'appréciation moyenne contre 0,556 avec enrichissement sémantique du domaine source avec un seuil de corrélation de 0,7). Comme attendu, notre méthode se comporte mieux pour des corrélations fortes pour lesquelles on peut parler d'invariants.

4.3.3 Enrichissement du domaine cible (Rs vers Dc et Ds vers Dc)

Nous nous intéressons maintenant à l'évaluation de la qualité de la complétion lorsqu'un enrichissement du domaine source est effectué ou non et qu'un enrichissement du domaine cible est effectué. La matrice du domaine cible des descripteurs D_c sur les utilisateurs de test U_{test} est une matrice présentant à l'origine 43% de remplissage.

Les MAE et couverture obtenues en fonction du seuil de corrélation sont présentées dans la figure 3. Les courbes continues représentent un transfert des ressources R_s vers les descripteurs D_c. Les courbes pointillées décrivent un transfert des descripteurs D_s vers les descripteurs D_c.

Tout d'abord nous pouvons constater que, que l'on enrichisse (D_s vers D_c) ou non (R_s vers D_c) le domaine source, la MAE est similaire, quelle que soit la valeur du seuil. Cependant, la couverture est plus élevée lorsque le domaine source est enrichi. Par exemple, lorsque le seuil de corrélation est de 0,5, la couverture est augmentée de 10 points de pourcentage pour D_s vers D_c par rapport à R_s vers D_c. D'autre part, dès que le seuil atteint 0,5, la couverture est comparable à celle des modèles originaux (43% à l'origine, 38,8% pour notre modèle). Si nous considérons comme invariants les couples corrélés à plus de 0,8, il y a 1114 couples (environ 0,01% du nombre total). Parmi ces couples, nous pouvons citer (*Tony_Curtis*, *film_culte*).

Nous en déduisons qu'enrichir sémantiquement le domaine source augmente la couverture de la complétion du domaine cible enrichi, tout en conservant une MAE équivalente.

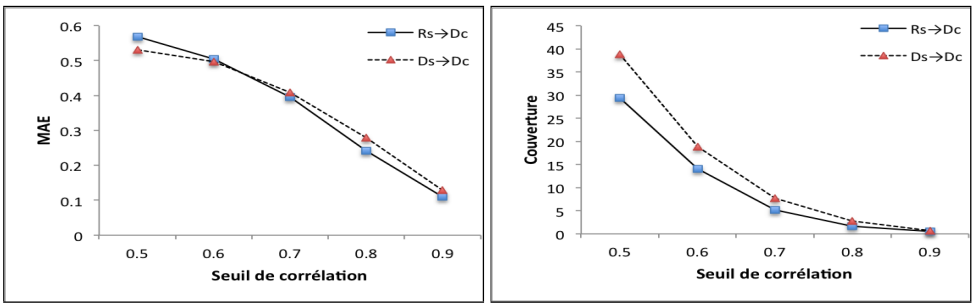

FIG. 3 – *MAE et couverture de la complétion du domaine cible enrichi*

D'une façon générale, nous pouvons conclure que l'exploitation de descripteurs sémantiques dans le domaine source mène à une MAE équivalente lorsque la complétion se fait sur le domaine cible enrichi tout en augmentant la couverture. Notre objectif est donc atteint.

5 Conclusion et perspectives

Dans cet article, qui est un premier travail, nous nous sommes intéressés au problème du manque de données dans le cadre de la modélisation utilisateurs. Partant du constat que, avec

le Web 2.0, les utilisateurs sont modélisés dans un nombre croissant de systèmes, nous avons exploré la possibilité de compléter les modèles utilisateur dans un système en exploitant les informations disponibles sur ces utilisateurs dans d'autres systèmes. Cette tâche est appelée médiation de modèles utilisateurs. Nous nous sommes intéressés plus particulièrement à la médiation inter-domaines de modèles utilisateurs : comment exploiter des informations provenant d'autres domaines, dits domaines sources, de façon à compléter un modèle utilisateur dans un autre domaine, dit domaine cible.

L'approche de médiation que nous proposons repose sur un modèle exploitant des couples corrélés entre domaines : soit des couples de ressources soit des couples de descripteurs sémantiques. Ces couples sont exploités de façon à compléter les modèles utilisateurs du domaine cible. Notre modèle de médiation a l'avantage d'être peu complexe et facilement compréhensible. Nous avons mené plusieurs expérimentations sur un corpus de données d'appréciations sur des films. Deux domaines ont été extraits, qui représentent chacun un genre de film. Ces expérimentations ont montré que l'exploitation d'informations sémantiques du domaine source permettaient d'augmenter la couverture de la complétion tout en conservant une MAE faible lorsque le domaine cible est enrichi, ce qui était notre objectif.

Ce travail préliminaire nous ouvre un grand nombre de perspectives. Nous envisageons dans un premier temps de réaliser plusieurs sélections aléatoires d'ensemble d'utilisateurs de test et de construire des corpus de données sur des domaines différents (nous pouvons par exemple imaginer le domaine des livres et celui de l'alimentaire) de façon à évaluer notre approche sur deux domaines distincts. Nous envisageons également de nous intéresser à la modélisation de liens entre domaines en exploitant des relations d'ordre plus grand que des relations descripteur à descripteur, en exploitant par exemple des règles d'association où l'antécédent serait composé de descripteurs du domaine source et le conséquent de descripteurs du domaine cible. Par la suite, nous chercherons à étudier la façon de faire le transfert entre domaines dans le cas où les utilisateurs ne sont pas communs entre les domaines, ou de façon plus générale, lorsqu'il n'est pas possible de faire le lien entre les utilisateurs de différents domaines. Dans ce cas, nous envisageons d'exploiter une classification d'utilisateurs.

Références

Adomavicius, G., Z. Huang, et A. Tuzhilin (2008). Personalization and recommender systems (book chapter). *Tutorials in Operations Research 2008 : State-of-the-Art Decision-Making Tools in the Information-Intensive Age, INFORMS*.

Adomavicius, G. et A. Tuzhilin (2005). Personalization technologies : a process-oriented perspective. *Communications of the ACM 48*, 83–90.

Berkovsky, S. (2006). Decentralized mediation of user models for a better personalization. In *Proc. of Int. Conf. on Adaptive Hypermedia and Adaptive Web-Based Systems, AH'06*, pp. 404–408.

Berkovsky, S., T. Kuflik, et F. Ricci (2007). Mediation of user models for enhanced personalization in recommender systems. *UMUAI 18*(3), 245–286.

Breese, J., D. Heckerman, et C. Kadie (1998). Empirical analysis of predictive algorithms for collaborative filtering. In *Proc. of UAI-98*.

Hofmann, T. et J. Puzicha (1999). Latent class models for collaborative filtering. In *Proc. of the sixteenth international joint conference on artificial intelligence*, pp. 688–693.

Jameson, A. (1999). User-adaptive systems : An integrative overview. In *UM'99, 7th International Conference on User Modeling*.

Koren, Y. (2008). Factorization meets the neighborhood : a multifaceted collaborative filtering model. In *Proc. of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 426–434.

Kostadinov, D. (2006). *Data Personalization : an approach for profile management and query reformulation*. Ph. D. thesis, Université de Versailles Saint-Quentin-en Yvelines.

Lam, S. et J. Riedl (2005). Privacy, shilling, and the value of information in recommender systems. In *Proc. of User Modeling Workshop on Privacy-Enhanced Personalization*, pp. 85–92.

Li, B., Q. Yang, et X. Xue (2009). Can movies and books collaborate ? : cross-domain collaborative filtering for sparsity reduction. In *Proc. of the 21st international jont conference on Artifical intelligence*, pp. 2052–2057. Morgan Kaufmann Publishers Inc.

Pan, W., E. W. Xiang, N. N. Liu, et Q. Yang (2010). Transfer learning in collaborative filtering for sparsity reduction. In *Association for the Advancement of Artificial Intelligence 2010*.

Shi, Y., M. Larson, et A. Hanjalic (2011). Tags as bridges between domains : Improving recommendation with tag-induced cross-domain collaborative filtering. In *UMAP '11*, pp. 305–316.

Zhang, Y., B. Cao, et D. yan Yeung (2010). Multi-domain collaborative filtering. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 725–732.

Zhou, J. et T. Luo (2009). Towards an introduction to collaborative filtering. In *Proc. of the 2009 International Conference on Computational Science and Engineering - Volume 04*, Washington, DC, USA, pp. 576–581. IEEE Computer Society.

Summary

Personalization services, which have emerged with Web 2.0, rely on the exploitation of user models. The more significant the amount of available information about users, the better the quality of the model and of the service. However, many services suffer from the data sparsity problem. In this paper, we choose to ease this problem by performing a cross-domain model mediation which relies on an information transfer using invariants or highly correlated pairs between domains. These pairs can be made up of resources or of semantic descriptors (after the user models have been semantically enhanced). We first show that the use of pairs of resources allows to fill missing values with new ones having a good intrinsic quality. Secondly, we show that the coverage of the users models which have been filled using semantic descriptors is larger with an equivalent quality. Therefore semantic enhancement is valuable for cross-domain transfer.