Le filtrage collaboratif et le web 2.0

Etat de l'art

Houda Oufaida — Omar Nouali

Division Théorie et Ingénierie des Systèmes Informatiques/DTISI
CEntre de Recherche sur l'Information Scientifique et Technique/CERIST
Rue des 3 frères Aïssou Ben Aknoun
BP 143 Alger 16030

{houfaida, onouali}@mail.cerist.dz

RÉSUMÉ. *Le présent article fait le point sur l'état de l'art des systèmes de filtrage d'information. Il présente les différentes techniques de filtrage proposées dans la littérature dont le filtrage à base de contenu, le filtrage collaboratif et les modèles de filtrage hybrides. Il présente également les différentes limitations dont souffre toujours ce type de systèmes notamment la rareté des votes et le problème du démarrage à froid. Enfin, il discute les évolutions majeures prévisibles avec l'introduction des aspects du web 2.0 tels que les ontologies, les réseaux sociaux et les annotations.*

ABSTRACT. *This paper overviews the recommender systems' field. It first presents different recommendation methods and techniques proposed in the literature from the content based recommender systems to the collaborative and hybrid recommendation models. It also describes the main limitations of the existing techniques such as the sparsity and the cold start problems. Finally, the new generation of semantic and social recommender systems are discussed which include several web 2.0 based extensions to overcome such weaknesses.*

MOTS-CLÉS : *filtrage d'information, web 2.0, web sémantique, réseau social, annotations.*

KEYWORDS: *recommender systems, web 2.0, semantic web, social networks, tagging.*

DOI:10.3166/DN.11.1-2.13-35

DN – 11/2008. Documents et web 2.0, pages 13 à 35

1. Introduction

De nos jours, d'importantes quantités d'informations sont à la disposition de chacun grâce au développement des technologies de l'information, le web en est un parfait exemple. De ce fait, le problème de la surcharge d'informations s'est très vite posé et constitue de nos jours un défi à surmonter.

Les systèmes de filtrage d'information s'inscrivent parmi ceux permettant la réception de documents jugés intéressants. Par opposition aux moteurs de recherche d'information (Google, AltaVista, Yahoo, etc.), qui requièrent de l'utilisateur une activité de formulation systématique de son besoin en utilisant des mots-clés. Le résultat retourné à l'utilisateur contient souvent un grand nombre de documents non pertinents. L'utilisateur doit donc sélectionner manuellement les documents pertinents. Il s'agit d'une tâche pénible et fastidieuse. Les systèmes de filtrage d'information pérennisent ce besoin d'information et permettent l'acheminement au cours du temps des documents intéressants.

Les premiers systèmes de filtrage d'information ont émergé à partir du début des années 1990, les expérimentations de (Foltz *et al.*, 1992) sur quatre méthodes de filtrage d'informations ont montré des résultats très prometteurs, (Resnick *et al.*, 1994) ont ensuite proposé leur célèbre système de recommandations de films, Grouplens[1] et depuis plusieurs travaux ont été menés et le filtrage d'information est devenu un axe de recherche très actif. Plusieurs applications ont été développées telles que des systèmes de recommandation de livres, de CDs ou autres par Amazon.com, de films avec Movielens[2], etc.

Plusieurs techniques ont été développées et un ensemble de modèles ont été proposés dans la littérature. Cet article fait le point sur l'état de l'art du filtrage d'information en général et plus précisément le filtrage d'information collaboratif.

L'introduction des aspects du web 2.0 dans de tels systèmes a été au centre des discussions ces cinq dernières années produisant une nouvelle génération de systèmes de filtrage d'information boostés par la sémantique ou les aspects sociaux du web sémantique, c'est ce qu'on appelle les systèmes de filtrage d'information *sémantique* ou *social*.

Cet article est organisé comme suit, la section 2 présente le filtrage d'information de manière générale, les différentes classifications possibles de ces systèmes et de manière plus détaillée le filtrage d'information collaboratif, la section 3 discute l'introduction des aspects du web 2.0 dans ce type de systèmes. Enfin, une approche initiale pour l'extension de ces systèmes est proposée.

1. Http://www.grouplens.org/
2. Http://www.movielens.org/

2. Le filtrage d'information

Le « Filtrage d'information » est l'expression utilisée pour décrire une variété de processus ayant pour but de fournir des informations à des personnes, informations en adéquation avec des centres d'intérêt de ces personnes (Belkin *et al.*, 1992).

Le filtrage peut être vu comme la sélection d'informations pertinentes sur un flux entrant. Le système fait une « prédiction » quant à l'intérêt que présente l'information pour l'utilisateur. Cette prédiction s'appuie sur le « profil » de cet utilisateur et aboutit à une prise de décision : « recommander » ou « ne pas recommander » l'information (Lopez, 2005).

Les profils, et parfois aussi la fonction de prédiction, évoluent dans le temps, à partir des informations cumulées et issues des documents déjà traités, de façon à ce que le profil traduise en permanence le besoin d'information de l'utilisateur.

Le problème de filtrage d'informations peut être formulé de la manière suivante (Adomavicius *et al.*, 2005) ; soit C un ensemble d'utilisateurs et S un ensemble de documents à recommander. Les deux ensembles peuvent être très grands contenant souvent des milliers de documents (utilisateurs) parfois même des millions dans certaines applications. Soit u une fonction qui mesure l'utilité que représentera un document s à un utilisateur c. On cherche alors des documents s' de manière à maximiser la fonction d'utilité u. D'une manière plus formelle on peut écrire :

$$U : C \times S \rightarrow R$$

$$\forall c \in C, s'_c = arg_{s \in S} max\ u(c,s)$$

L'utilité d'un document est souvent représentée par un vote ou une note soit donnée par l'utilisateur de manière *explicite* soit estimée par le système de manière *implicite*. Chaque utilisateur c de l'espace C est représenté par un *profil*, ce profil peut ne contenir que les votes de cet utilisateur dans les cas les plus simples, et peut être aussi plus complet contenant d'autres informations sur l'utilisateur, démographiques par exemple (sexe, âge, profession, situation familiale…).

Traditionnellement, les systèmes de filtrage d'informations ont été classés en trois catégories : les systèmes à base de contenu, les systèmes de filtrage collaboratif (voir la section 2.2) et les systèmes de filtrage hybrides. Cette classification dépend de la manière avec laquelle l'utilité ou la pertinence éventuelle est calculée ou estimée.

2.1. *Les systèmes de filtrage à base de contenu*

Les systèmes de filtrage à base de contenu recommandent des documents similaires à ceux que l'utilisateur a déjà apprécié. Ceci est calculé en rapprochant les centres d'intérêt des utilisateurs (introduits de manière explicite à travers un questionnaire par exemple ou de manière implicite à travers la surveillance de son

comportement) avec la métadonnée ou les caractéristiques des documents, sans prendre en compte les avis des autres utilisateurs (Peis *et al.*, 2008).

Deux fonctionnalités centrales ressortent de ce type de systèmes :

– la sélection des documents pertinents vis-à-vis du profil ;

– la mise à jour du profil en fonction du retour de pertinence fourni par l'utilisateur sur les documents qu'il a reçus.

Par exemple, avec cette approche, un système de recommandation de films *essaie* de détecter le maximum d'attributs en communs entre deux films, le premier bien apprécié par l'utilisateur actif dans le passé et le nouveau film (le même genre, même directeur, même acteur…).

Le contenu du profil dépend de la méthode utilisée dans l'analyse du contenu des documents. Pour ce faire, plusieurs techniques d'apprentissage en ligne ont été utilisées ; des techniques d'analyse textuelle empruntées à la recherche d'informations pour la recommandation de documents textuels (sites web, articles…) (Balabanovic *et al.*, 1997 ; Pazzani *et al.*, 1997 ; Pazzani, 1999). Le profil est souvent sous forme d'un vecteur de mots-clés avec des poids [1]. Le poids associé à chaque mot reflète l'importance de ce terme pour l'utilisateur. Ces mots sont souvent extraits à l'aide de la mesure *TF-IDF* (Salton, 1989). Ce vecteur est ensuite comparé à celui du document [2], pour ce faire plusieurs mesures peuvent être utilisées telles que la mesure des vecteurs cousine [3].

$$Profile\_contenu\ (c)=\{(t^c_j,\ w^c_j)\},\ j=1\ldots k \qquad [1]$$

$$Contenu\ (s)=\{(t^s_i,\ w^s_i)\},\ i=1\ldots n \qquad [2]$$

$$u(c,s) = \cos(\vec{w}_C, \vec{w}_S) = \sum_{i=1,k} \frac{w_{i,c}}{\sqrt{\sum_{i=1,k} w_{i,c}^2}} \frac{w_{i,s}}{\sqrt{\sum_{i=1,k} w_{i,s}^2}} \qquad [3]$$

Avec :

$w_{i,c}$ Le poids du terme *i* dans le vecteur du profil utilisateur

$w_{i,s}$ Le poids du terme *i* dans le vecteur du document *s*

En plus de ces méthodes heuristiques, d'autres systèmes à base de modèles ont été proposés dans la littérature ; tels que les classifieurs bayesiens (Mooney *et al.*, 1998 ; Pazzani *et al.*, 1997), les réseaux de neurones (Pazzani *et al.*, 1997), les arbres de décision (Zhang *et al.*, 2002).

Ce type de filtrage souffre d'un certain nombre de limitations, on peut citer son incapacité de recommander les documents non textuels qui ne disposent pas d'informations sur leurs contenus (multimédia par exemple). En plus, les critères qualité et fiabilité de la source ne sont pas considérés (Adomavicius *et al.*, 2005).

L'effet *entonnoir* restreint le champ de vision des utilisateurs ; ce type de filtrage est incapable de recommander des documents qui sont différents de ceux que l'utilisateur a déjà vu et évalué. L'utilisateur n'aura donc jamais l'occasion de voir et juger ces nouveaux documents différents mais peut être intéressants. De plus, l'utilisateur doit évaluer un nombre suffisant de documents pour que le système puisse lui recommander des documents pertinents, ceci n'est pas le cas pour les nouveaux utilisateurs. Ce problème est connu dans la littérature comme celui du nouvel utilisateur.

2.2. *Les systèmes de filtrage collaboratif*

Les systèmes de filtrage d'information collaboratif ont été largement investis et surtout adoptés dans différentes applications. Parmi les premières applications on peut retrouver Tapestry (Golberg *et al.*, 1992), Grouplens (Resnick *et al.*, 1994), Ringo (Shardanand *et al.*, 1995). Parmi les applications commerciales, on peut aussi citer Amazon (Linden *et al.*, 2003), netflix[3] , Barnes & Noble[4] etc.

Contrairement aux systèmes de filtrage à base de contenu, les systèmes de filtrage collaboratif ont pour principe d'exploiter les « évaluations » que des utilisateurs ont faites de certains documents, afin de recommander ces mêmes documents à d'autres utilisateurs. De manière plus formelle, l'utilité d'un document s pour un utilisateur c, $u(c,s)$ sera calculée en fonction des $u_j\ (c_j,s)$ qui lui sont *similaires*. Ainsi, la fonction de prédiction F utilise la matrice des votes $C \times S \rightarrow [1,10]$ (voir tableau 1) et procède en deux étapes (Woerndl *et al.*, 2007) :

– calculer la similarité entre utilisateurs et inférer les communautés ;

– prédire des notes pour quelques documents et ne sélectionner que les documents avec un score élevé.

Tout le problème est donc de déterminer l'ensemble d'utilisateurs proches de l'utilisateur actif sur la base des jugements donnés en commun c'est-à-dire portés sur les mêmes documents.

	S_1	S_2	S_3	S_4	S_5
C_1			7	6	
C_2			5	6	7
C_3			6	6	7
C_4	7	5			7
C_5	7	6			7

Tableau 1. *La matrice Utilisateur/Document*

3. Http://www.netflix.com/
4. Http://www.barnesandnoble.com/

Il existe deux grandes approches collaboratives, une approche basée *mémoire* et une autre basée *modèle*. Le choix de l'approche à utiliser dépend des informations prises en compte lors du calcul de la prédiction.

2.2.1. *Filtrage basé mémoire*

Les algorithmes basés mémoire (Resnick *et al.,* 1994 ; Nakamura *et al.*, 1998 ; Delgado *et al.*, 1999), utilisent la *totalité* ou une *partie* des profils utilisateurs afin de générer une nouvelle prédiction. Ainsi la note potentielle que donnera un utilisateur à un document est calculée en fonction [4] des notes données par les autre utilisateurs (généralement les N plus proches) au même document.

$$p_{c,s} = \bar{v}_c + k \sum_{i=1}^{n} w(c,c_i)(v_{i,s} - \bar{v}_i) \qquad [4]$$

Avec :

$v_{i,s}$ L'évaluation du document *s* par l'utilisateur c_i

$\bar{v}_i$ La moyenne des évaluations fournies par l'utilisateur c_i

$$\bar{v}_i = \frac{1}{|S_i|} \sum_{j \in S_i} v_{i,j}$$

$w(c,c_i)$ Le score de similarité entre l'utilisateur c_i et c

k Un coefficient de normalisation

n Le nombre d'utilisateurs considérés.

Le score de similarité peut être calculé selon différentes formules, une des plus utilisées est celle du coefficient de Pearson [5] (Resnick *et al*, 1994 ; Shardanand *et al.*, 1995) ou la méthode du vecteur cousine [6] (Breese *et al.*, 1998 ; Sarwar *et al.*, 2001).

$$w(c,c_i) = \frac{\sum_{j \in S_{c,c_i}} (v_{c,j} - \bar{v}_c)(v_{i,j} - \bar{v}_i)}{\sqrt{\sum_{j \in S_{c,c_i}} (v_{c,j} - \bar{v}_c)^2 \sum_{j \in S_{c,c_i}} (v_{i,j} - \bar{v}_i)^2}} \qquad [5]$$

$$w(c,c_i) = \sum_{j \in S_{c,c_i}} \frac{v_{c,j}}{\sqrt{\sum_{k \in S_c} v_{c,k}^2}} \frac{v_{i,j}}{\sqrt{\sum_{k \in S_{c_i}} v_{i,k}^2}} \qquad [6]$$

Avec :

$v_{c,j}$ L'évaluation du document *j* par l'utilisateur actif *c*

$\bar{v}c$ La moyenne des évaluations données par l'utilisateur actif *c*

$v_{i,j}$ L'évaluation donnée par l'utilisateur c_i

$\bar{v}i$ La moyenne des évaluations de l'utilisateur c_i

S_c, S_{c_i} L'ensemble des documents évalués par l'utilisateur *c* (Resp. c_i)

S_{c,c_i} L'ensemble des documents évalués à la fois par *c* et c_i

De tels algorithmes ont l'avantage d'être simples à mettre en œuvre et d'évoluer dynamiquement en fonction des profils utilisateurs. En effet, toute évolution d'un utilisateur se répercute directement dans le calcul de prédiction.

Cependant, ces algorithmes souffrent de deux inconvénients majeurs. D'une part, la forte complexité combinatoire empêche le passage à l'échelle pour un nombre important d'utilisateurs et de ressources. D'autre part, le faible nombre de ressources communément évaluées par les utilisateurs engendre des prédictions peu pertinentes (Lumineau, 2002).

2.2.2. *Filtrage basé modèle*

Le filtrage basé modèle *apprend* un *modèle* descriptif liant les utilisateurs, les documents et les votes. D'un point de vue probabiliste, le processus de filtrage prédit la valeur d'un vote donné $v_{c,s}$ compte tenu du profil utilisateur ou de ses précédents votes $v_{c,k}$ [7]

$$P_{c,s} = E(v_{c,s}) = \sum_{i=0}^{m} \Pr(v_{c,s} = i \mid v_{c,k}, k \in S_c) \cdot i \qquad [7]$$

Pour estimer cette probabilité, (Breese *et al.*, 1998) ont proposé l'utilisation de deux modèles : un modèle de clusters et un modèle réseaux bayesiens.

Le modèle à base de clusters repose sur le principe que certains groupes ou types d'utilisateurs capturent un ensemble commun de préférences et de goûts. Du point de vue formel, on s'appuie sur un classifieur bayésien, le modèle de probabilité qui met en relation les probabilités jointes des classes et des évaluations, avec l'ensemble de distributions conditionnelles et marginales, ce qui représente la formulation standard « naïve » de Bayes [8]. Ici, le nombre de classes et les paramètres du système sont calculés à partir du jeu de données, l'algorithme EM (Dempster *et al.*, 1977) est utilisé.

$$\Pr(C = c, v_1, ..., v_n) = \Pr(C = c) \prod_{i=1}^{n} \Pr(v_i \mid C = c) \quad [8]$$

Dans le modèle à base de réseaux bayesiens proposé par (Breese *et al*., 1998), les nœuds correspondent aux documents. Les états pour chaque nœud correspondent aux valeurs d'évaluation possibles. Il inclut également un état correspondant à l'absence d'évaluation pour les domaines où il n'y a pas d'interprétation naturelle des données manquantes. Le réseau est ainsi construit à partir des données en appliquant un algorithme d'apprentissage (Chickering *et al*., 1997), le résultat est alors sous forme d'*arbres de décision* représentant chaque table de probabilité conditionnelle pour chaque nœud.

D'autres chercheurs ont utilisé des techniques d'apprentissages pour construire le modèle. (Billsus *et al.*, 1998) ont traité le processus de filtrage collaboratif comme un problème de classification couplé avec une méthode d'analyse de données (la SVD *Singular Value Decomposition*). En outre, (Hoffman *et al.*, 2004) proposent l'utilisation de l'analyse par la sémantique latente pour la construction des communautés utilisateurs.

Les approches basées modèle ont l'avantage de mieux traiter le problème du manque d'évaluations ceci en regroupant les utilisateurs et les documents en groupes ou classes. Cependant, pour une large base de données ces modèles deviennent non pratiques (Yu *et al.*, 2004).

Le grand avantage avec les méthodes collaboratives (à base de mémoire ou de modèle) est qu'elles peuvent être appliquées à tout type de documents ; textuels, multimédias et sont ainsi, contrairement aux méthodes à base de contenu, totalement indépendantes du format du document à recommander. En plus, on est capable de diffuser des documents non nécessairement similaires à ceux déjà reçus.

Cependant, le nombre d'évaluations déjà attribuées par les utilisateurs est toujours très inferieur par rapport à celui qu'on doit prédire (1 % pour le jeu de données MovieLens et 3 % pour Netflix). En plus, parmi ces évaluations on compte très peu d'évaluations en commun entre deux utilisateurs ce qui conduit à des scores de similarités peu pertinents.

Un autre problème dont souffrent les systèmes de filtrage collaboratif est le *démarrage à froid* ; où ils sont incapables de produire des recommandations pour les nouveaux utilisateurs du fait qu'ils n'ont pas encore donné d'évaluations, leurs communautés par l'historique des évaluations sont toujours inconnues. Du fait que les documents à recommander ne sont décrits que par les évaluations fournies par les utilisateurs, les nouveaux documents ne peuvent être recommandés. La combinaison des deux situations (nouveaux utilisateurs et les nouveaux documents) conduit à une situation de démarrage à froid pour un nouveau système où les performances sont très mauvaises en raison de l'absence d'informations sur lesquelles fonder le processus de filtrage personnalisé.

Ce problème est généralement traité en combinant les méthodes purement collaboratives avec celles basées contenu pour mieux décrire les nouveaux documents ou en utilisant des sources de données externes susceptibles d'alimenter le système telles que les données sociodémographiques comme l'âge, la profession, etc. ou même celles permettant d'estimer les préférences des utilisateurs sur certains attributs des documents à recommander, comme par exemple les acteurs préférés lorsqu'il s'agit de recommander des films. Ceci donne lieu à de nouvelles approches, dites *hybrides*.

2.3. *Les systèmes de filtrage hybride*

Les méthodes hybrides cherchent à atténuer les insuffisances de chacune des deux précédentes approches en les combinant de différentes manières.

(Claypool *et al.*, 1999) ont combiné les recommandations produites par les deux méthodes à base de contenu et collaboratives appliquées séparément. (Pazzani, 1999) a appliqué les algorithmes de filtrage collaboratifs sur une matrice décrivant les préférences des utilisateurs sous forme de mots-clés pondérés au lieu de la traditionnelle matrice des votes. (Mellville *et al*., 2002) ont utilisé les recommandations à base de contenu pour compléter la matrice des votes et ensuite appliquer l'algorithme de recommandation collaboratif sur cette matrice.

Plusieurs autres travaux se sont orientés vers la construction de modèles unificateurs qui prennent en compte les caractéristiques des utilisateurs et des documents ; (Basu *et al.*, 1998) ont proposé l'application d'un classifieur d' attributs sur les utilisateurs couplé avec des informations sur le contenu des documents (ici le genre, les acteurs, le directeur… d'un film). Ainsi, il est possible de recommander de nouveaux documents sur la base des préférences des utilisateurs vis-à-vis de ces attributs sans disposer d'aucune évaluation au départ (Hauger *et al.*, 2007). (Shein *et al.*, 2002) proposent plusieurs modèles de probabilités où l'idée de base est l'analyse par la sémantique latente pour identifier de possibles liens sémantiques (une affinité particulière par exemple) cachés liant un document et un utilisateur. De plus, (Nguyen *et al.*, 2007) utilisent les données disponibles à froid recueillies à partir de l'utilisateur dès son inscription, son âge, sa profession, son lieu de résidence etc. Ces données sont utilisées pour compenser partiellement les données non fournies initialement par les utilisateurs afin de les positionner au sein de leurs communautés initiales.

Récemment, une nouvelle génération de systèmes de recommandations boostés par les aspects du web sémantique ont émergé. Ces systèmes exploitent des outils tels que les taxonomies, les ontologies, les réseaux sociaux et les annotations. La prochaine section présente ces nouvelles issues.

3. Le filtrage collaboratif et le web 2.0

La plupart des techniques de filtrage actuelles ne prennent pas en compte un bon nombre d'informations utiles pour *expliquer* un jugement qu'effectue un utilisateur. Ceci requière la prise en compte de différentes sources d'informations sur utilisateur (ses préférences précédemment annoncées, son entourage social) et leurs relations avec le contenu des documents qui lui sont proposés.

Par exemple, dans les systèmes purement collaboratifs, l'utilisateur et le document sont représentés par un ensemble de votes. Par conséquent, le groupement de ces utilisateurs dans des communautés traduit une similarité *globale* souvent insuffisante pour refléter les relations pouvant lier ces utilisateurs et encore plus les documents (Golbeck, 2008).

3.1. *Systèmes de filtrage d'information à base d'ontologies*

Avec l'émergence du web sémantique, la disposition de larges taxonomies sur le web a encouragé leurs utilisations pour la classification des documents et produits. Par exemple, la taxonomie UNSPSC[5] fournit une terminologie pour les produits et les services contenant plus de 22 000 termes, la classification hiérarchique des livres Amazon.com[6] contient plus de 11 000 codes, l'Open Directory Project (ODP[7]) contient plus de 4 578 875 sites catégorisés en 590 000 catégories, etc., Ces taxonomies sont simplement considérées comme une source de connaissance sémantique sur le domaine d'application par rapport au système de filtrage.

L'utilisation de ressources conceptuelles pour indexer les documents a fait émerger de nouvelles approches tentant d'intégrer la sémantique dans la recherche d'informations (Baziz *et al.*, 2005 ; Khan *et al.,* 2002). Ces méthodes ont inspiré le passage d'une description des documents par mots-clés ou à base d'attributs à une description sémantique à base de concepts.

Dans les systèmes de filtrage d'informations, ces ressources sémantiques sont surtout utilisées pour classifier les documents et inférer les profils utilisateurs. Quickstep (Middelton *et al.*, 2004) est un système de filtrage d'articles scientifiques hybride. Il utilise une technique de classification supervisée couplée avec une représentation ontologique des domaines de recherche afin d'extraire les centres d'intérêts de l'utilisateur. Il assigne chaque article à une classe (thème) avec laquelle le vecteur représentatif du document est le plus similaire. Le profil lui-même est calculé à partir de la corrélation entre les articles explorés et les thèmes abordés dans ces articles. En résultat, le profil sera constitué d'un ensemble de thèmes extraits

5. Http://www.unspsc.org/
6. Http://www.amazon.com/
7. Http://www.dmoz.org/

d'une ontologie avec leurs poids. Les thèmes les plus généraux sont aussitôt inférés grâce à la hiérarchie de concepts présente dans l'ontologie.

(Mobasher *et al.,* 2004) proposent une mesure de similarité qui combine la similarité sémantique entre documents avec celle basé sur les votes [9].

$$CombinedSim(i_p,i_q)=\alpha \,.\, SemSim(i_p,i_q)+(1-\alpha) \,.\, RateSim(i_p,i_q) \quad [9]$$

Avec :

α Un paramètre de combinaison sémantique spécifiant le poids de la similarité sémantique dans la combinaison linéaire.

La similarité sémantique est calculée sur la base d'une description sémantique structurée des documents automatiquement extraite d'une base de données en ligne sur le cinéma mondial[8] à l'aide d'une ontologie de domaine (figure 1). L'union de ces descriptions forme une matrice, une analyse par la sémantique latente est ensuite appliquée pour réduire sa dimension.

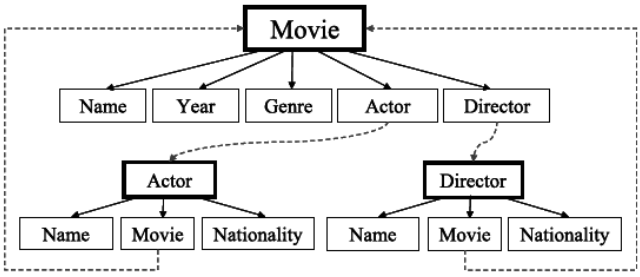

Figure 1. *Une partie de la représentation sémantique d'un film*

Les expérimentations ont montré que cette méthode était particulièrement bénéfique lorsque la matrice des votes était creuse ou dans le cas du nouveau document.

Dans le même contexte, (Zieigler *et al.*, 2004) utilisent une taxonomie décrivant les documents à recommander pour inférer les profils utilisateurs. Le profil est ainsi constitué de *catégories* de documents avec un score d'intérêt plutôt que d'*instances* de ces catégories. Ceci est utilisé pour découvrir d'autres profils similaires.

(Zuber *et al.*, 2006) ont aussi proposé l'utilisation des ontologies pour compléter le modèle utilisateur et améliorer la pertinence des recommandations même avec un nombre minimal de votes. (Lops *et al.*, 2007) proposent un système de filtrage hybride ; chaque profil utilisateur est constitué de deux vecteurs de concepts

8. Http://www.imdb.com/

Wordnet[9] (t^m_i), pondérés (w^m_i). Le premier représente les centres d'intérêts positifs (évaluation positive) [10], le second représente ceux négatifs (évaluation négative) [11]. Les communautés sont formées à l'aide de la méthode des k-moyennes (MacQueen, 1967). L'union des clusters positifs et négatifs forme le voisinage de l'utilisateur.

$$P^+_u=\{(t^m_i, w^m_i)\},\ m=1...5,\ i=1...n \quad [10]$$

$$P^-_u=\{(t^m_j, w^m_j)\},\ m=1...5,\ j=1...k \quad [11]$$

(Mehta, 2005) définit une description d'un modèle utilisateur *unificateur* et *extensible* UUCM (*Unified User Context Model*) en vue de permettre l'échange des profils utilisateurs entre systèmes de filtrage d'information ou de personnalisation de manière plus générale.

3.2. *Systèmes d'annotations collaboratives*

Les systèmes d'annotations collaboratives sont un autre aspect très prometteur du web sémantique. Ces systèmes offrent à leurs utilisateurs la possibilité d'héberger leurs photos, vidéos, documents ou toute autre ressource et surtout leur assigner un ensemble de mots décrivant leurs contenu : c'est ce qu'on appelle les *annotations* ou *tags*. Ces annotations décrivent le contenu de la ressource ou bien fournit une information contextuelle et sémantique en plus. Parmi les services web les plus populaires on retrouve Flickr[10], del.icio.us[11]. En général, les annotations sont souvent associées au web sémantique du fait qu'ils permettent aux utilisateurs d'ajouter une métadonnée aux documents d'une manière très simple.

Cette connaissance en plus d'offrir les annotations sur les documents peut être exploitée dans le processus de filtrage d'informations. Des travaux très récents ont investi cette issue.

(Karen *et al.,* 2008) proposent d'étendre la matrice des votes avec les annotations comme troisième dimension.

Cette matrice 3-dimensionnel est ensuite décomposée en trois sous matrices : <utilisateurs, documents>, <utilisateurs, annotations> et <documents, annotations>. Pour le calcul de la prédiction, les auteurs utilisent la combinaison de la similarité entre utilisateurs et celle entre documents mais cette fois sur la matrice étendue. Les résultats ont montré que l'introduction des annotations a été surtout bénéfique pour la combinaison des deux similarités ce qui n'était pas du tout le cas si on l'appliquait

9. Base de données lexicale développée par des linguistes du laboratoire des sciences cognitives de l'université de Princeton.
10. Http://flickr.com
11. Http://del.icio.us

à chaque méthode a part (similarité utilisateur/utilisateur et similarité document/document).

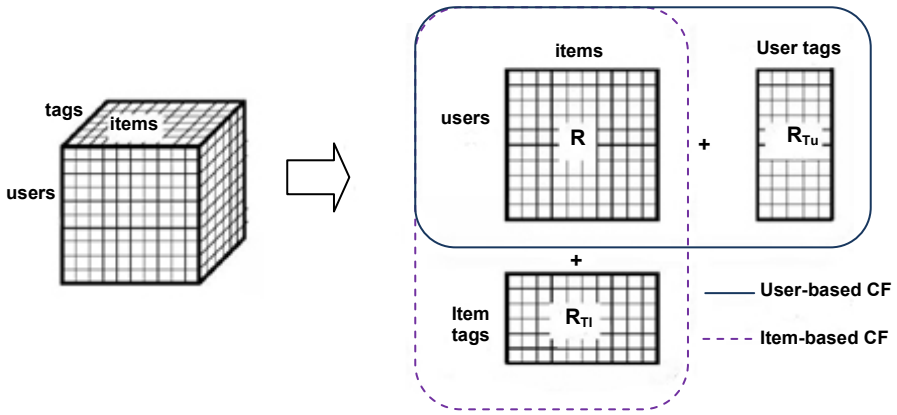

Figure 2. *Extension de la matrice utilisateur/document par les annotations*

(Szomszor *et al.*, 2007) proposent l'utilisation des annotations pour enrichir le profil utilisateur. Chaque utilisateur dispose d'un *nuage* d'annotations, les documents aussi. La recommandation est ainsi calculée en utilisant la similarité sémantique entre les deux nuages.

3.3. *Systèmes de filtrage d'informations à base de réseaux sociaux*

Le grand succès des services de réseautage social en ligne, où les utilisateurs peuvent gérer des listes d'amis et manifester leurs centres d'intérêts, a encouragé la réutilisation de ces données sociales dans les systèmes de filtrage d'information. Facebook[12] (70 millions d'utilisateurs), MySpace[13](110 millions d'utilisateurs) en sont des célèbres exemples.

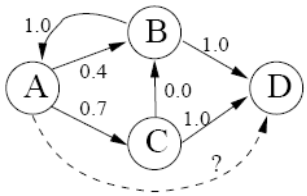

Figure 3. *Un exemple d'un réseau social avec les scores de confiance (Massa* et al.*, 2004)*

12. Http://www.facebook.com/
13. Http://www.myspace.com/

Dans les domaines subjectifs, pour qui la majorité des systèmes de filtrage d'information ont été conçus, les facteurs sociaux peuvent jouer un rôle important dans le processus de prise de décision. Ainsi, la question qui se pose est comment peut-on exploiter les modèles sociaux dans les systèmes de filtrage d'informations ?

Une première idée serait de remplacer la formation des communautés classique sur la base des votes avec celle induite par le réseau social (amis et amis d'amis). (Sinha *et al.*, 2001) ont comparé les recommandations collaboratives classiques avec celles faites par les amis sur trois systèmes de recommandations de films (Amazon.com, MovieCritic et Reel.com) et trois autres pour la recommandation de livres (Amazon.com, RatingZone et Sleeper). Les résultats ont montré que les utilisateurs ont préféré celles faites par leurs amis. Ceci peut être expliqué par le fait que les amis sont plus qualifiés à les conseiller puisqu'ils sont supposés connaitre davantage les préférences des utilisateurs.

(Groh *et al.*, 2007) ont mené une étude comparative de la performance d'un algorithme de filtrage collaboratif classique par rapport à un algorithme de filtrage social où la communauté sociale est constituée des amis de l'utilisateur [12] et de leurs amis [13].

$$F^{(u)}=\{u_j \mid A_{ij}=1\} \quad [12]$$

$$N_{social}^{(u)}= F^{(u)} \cup \{u_k \mid \exists\ u_j \in\ F^{(u)} : A_{jk}=1\} \quad [13]$$

Avec :

$A_{ij}=1$ Il existe un arc liant l'utilisateur i à j dans le réseau social (i et j sont amis)

La prédiction des votes est calculée en considérant l'union des deux communautés, collaborative et sociale. Après expérimentation, les auteurs ont constaté que l'approche sociale est nettement plus performante dans le cas ou on ne dispose pas de suffisamment de votes.

(Ziegler *et al.,* 2007) ont mené une étude expérimentale sur la relation entre la similarité des profils utilisateurs et le concept de *confiance sociale.* Ils ont développé un site web de réseautage social en ligne, FilmTrust[14], où les utilisateurs peuvent passer en vue des films, les évaluer et gérer leurs profils en utilisant le vocabulaire FOAF[15] (*Friend of a Friend vocabulary*). Les auteurs proposent un algorithme pour la propagation du degré de confiance sociale à travers le réseau social, TidalTrust [14].

14. Http://trust.mindswap.org/FilmTrust/
15. Http://www.foaf-project.org/

$$t_{i,s} = \frac{\sum_{j \in adj(j) / t_{i,j} \geq \max} t_{i,j} t_{j,s}}{\sum_{j \in adj(j) / t_{i,j} \geq \max} t_{i,j}} \quad [14]$$

Avec :

$t_{i,s}$	Le degré de confiance entre l'utilisateur actif *i* et la cible *s*
$adj(j)$	L'ensemble des utilisateurs adjacents à l'utilisateur *j*
$t_{i,j}$	Le degré de confiance liant *i* à *j*
max	Un seuil du degré de confiance

Pour la prédiction des votes, le système sélectionne les utilisateurs ayant une valeur de confiance maximale. L'algorithme considère, ainsi, la *moyenne* des votes donnés par ces utilisateurs, celle-ci étant plus *fiable*, pondérée par leurs scores de confiance [15].

$$r_{s,m} = \frac{\sum_{i \in S} t_{s,i} r_{i,m}}{\sum_{i \in S} t_{s,i}} \quad [15]$$

Avec :

$r_{s,m}$	La prédiction de l'évaluation du document *m* par l'utilisateur *s*
$r_{i,m}$	L'évaluation donnée par l'utilisateur *i* sur le document *m*

Les résultats ont montré que plus le degré de confiance est grand plus la différence des votes diminue indiquant ainsi une *forte* corrélation entre la similarité et la confiance sociale.

(Massa *et al.* 2004), présentent le modèle « Web of Trust » où les utilisateurs définissent un ensemble d'amis à qui ils font *confiance*. Ce modèle à comme entrée la matrice des votes <utilisateurs, documents> et la matrice des scores de confiances entre utilisateurs <utilisateurs, utilisateurs> et produit en sortie une matrice des votes estimés.

4. Synthèse des approches de filtrage d'information

Le tableau 2 synthétise les différentes méthodes de filtrage d'information. Ceci en termes de représentation du profil utilisateur adoptée, des sources de données et des techniques utilisées lors du calcul des recommandations ainsi que les avantages et inconvénients propres à chaque méthode.

	Profil utilisateur	**Sources de données**	**Techniques utilisées**	**Avantages (+) Inconvénients(-)**
Filtrage basé contenu	-Mots-clés	-Evaluations -Documents indexés	-Similarité vectorielle	(+) Nouveaux documents (-) Le facteur qualité non pris en compte (-) Effet entonnoir (-) Difficulté de recommander des documents non textuels (-) Nouvel utilisateur
Filtrage collaboratif basé mémoire	-Evaluations	-Evaluations explicites ou implicites	-Similarité vectorielle	(+) Indépendant du format des documents (+) Diversité (+) Qualité prise en compte (+) Evolution dynamique du profil (-) Démarrage à froid (-) Rareté des votes (-) Complexité
Filtrage collaboratif basé modèle	-Evaluations	-Evaluations explicites ou implicites	-Classification -Analyse par la sémantique latente	(+) Compense le manque des votes (+) Rapidité du calcul (-) Complexité de mise à jour
Filtrage hybride	-Evaluations + -Mots-clés	-Evaluations -Documents indexés	-Combinaison linéaire : collaboratif, basé contenu -Cascade : basé contenu puis collaboratif -Modèles unificateurs	(+) S'adapte mieux avec le problème de démarrage à froid (-) La performance dépend fortement du domaine
Filtrage basé ontologies	-Evaluations + -Concepts	- Evaluations -Ressources sémantiques	-Similarité sémantique -Similarité vectorielle	(+) Profil utilisateur plus complet (+) Meilleure précision (-) Connaissance sémantique exigée

	Profil utilisateur	**Sources de données**	**Techniques utilisées**	**Avantages (+) Inconvénients(-)**
Filtrage basé annotations	-Evaluations + -Annotations	-Evaluations -Annotations	-Similarité sémantique	(+) Les annotations n'exigent aucune connaissance particulière du domaine (-) Vocabulaire non contrôlé
Filtrage social	-Evaluations + -Entourage social	-Evaluations de l'utilisateur -Evaluations des amis	-Scores de confiance entre l'utilisateur et ses amis	(+) Disponibilité de larges communautés sociales sur le web (+) Compense les votes manquants (+) N'exige aucun effort cognitif (-) Non adapté aux domaines rationnels

Tableau 2. *Synthèse des approches de filtrage d'information*

5. Conclusion et perspectives

Les systèmes de filtrage d'information ont connu une avancée significative ces dix dernières années, depuis les premiers systèmes collaboratifs classiques à ceux à base de contenu ou hybrides. Ils ont été largement investis dans divers domaines tels que le commerce électronique (livres, cinéma, musique, voyages, restauration, etc.)

Toutefois, des problèmes subsistent toujours, parmi lesquels on peut citer le démarrage à froid et la rareté des votes. L'hybridation avec des méthodes à base de contenu a été essentiellement utilisée pour les résoudre.

Dans cet article, nous avons mis le point sur l'état de l'art des travaux sur le filtrage d'information en général et en particulier sur le filtrage d'information collaboratif. Nous avons aussi exposé des solutions basées web 2.0 pour traiter les problèmes liés à ce type de systèmes : génération des profils utilisateurs guidée par des taxonomies ou des ontologies, les recommandations sociales et l'utilisation des annotations pour compléter la description des documents. Les résultats des premières expérimentations ont été très prometteurs en termes de :

– l'exploitation de sources de données sur le domaine (les ontologies par exemple) enrichit la description des utilisateurs et des documents ;

– cette représentation boostée par la sémantique améliore la classification des documents et le groupement en communautés des utilisateurs, produisant ainsi de meilleures recommandations en termes de précision et de couverture ;

– la possibilité de l'utilisation de langages standard issus du web 2.0 pour ce type de système : le vocabulaire FOAF (*Friend Of A Friend*, ou « l'ami d'un ami ») pour décrire les personnes, leurs comptes, ce quelles font, les liens entre elles, l'appartenance à des groupes. Le standard APML (*Attention Profiling Mark-up Language*) qui permet de modéliser les préférences d'un utilisateur et leurs poids. L'initiative Data Portabilité, DP et les standards ouverts permettent de partager les données d'une manière plus libre et l'agrégation de profils utilisateurs entre différents sites pour pouvoir réutiliser les informations personnelles, les listes de contacts, les préférences etc. ;

– l'introduction des réseaux sociaux tels que les communautés sociales disponibles sur le web augmenterai la confiance des utilisateurs envers le système de recommandations et encouragerait les utilisateurs à mieux participer et par conséquent à fournir d'avantage d'informations sur leurs goûts et préférences

En intégrant toute source d'information disponible sur l'utilisateur, une réflexion possible serait de disposer d'un schéma (figure 4) où le profil utilisateur sera composé de trois dimensions :

– *Dimension évaluations* : contiendra l'ensemble des évaluations données par cet utilisateur, elles sont recueillies au fur et à mesure que l'utilisateur évalue des documents soit de manière explicite soit de manière implicite ;

– *Dimension sociale* : contiendra l'ensemble des données personnelles relatives à cet utilisateur : nom, prénom, date de naissance, sexe, profession, Email, l'ensemble des contacts d'ordre professionnel et ou personnel (liste d'amis), page web personnel, blog. Elles sont collectées dès son inscription et peuvent être mises à jour au cours du temps ;

– *Dimension sémantique* : contiendra les centres d'intérêts sous forme de concepts ou thèmes avec des poids reflétant leurs degrés d'importance vis-à-vis l'utilisateur. Ces concepts peuvent être extraits automatiquement par un processus d'indexation sémantique des documents évalués, cette indexation étant guidée par une ontologie de domaine par exemple.

La description des dimensions sociale et sémantique peut être faite en utilisant le vocabulaire FOAF. Chacune de ces dimensions est exploitée, comme le montre la figure 4, pour inférer des communautés collaboratives, sociales et sémantiques. La communauté collaborative classique se base sur la similarité des votes, la communauté sociale contiendra l'ensemble des amis de l'utilisateur et les amis de ses amis et enfin, la communauté sémantique, peut être calculée à partir des scores de similarité sémantique globale en tenant compte de tous les concepts de la dimension sémantique de chaque centre d'intérêt à part.

Chaque dimension produit un ensemble de recommandations, un classement de ces recommandations est alors nécessaire. Ce classement peut être *adaptatif* [16] : les recommandations sociales peuvent, par exemple, être prioritaires dans le cas où

on ne dispose pas ou de très peu d'évaluations et la dimension sémantique n'est pas encore *claire* (nouveau utilisateur) ou encore pour quelques catégories de documents (de loisir : musique par exemple). Les recommandations collaboratives peuvent être prioritaires si on veut, par exemple, découvrir de nouveaux centres d'intérêts chez l'utilisateur en lui proposant des documents qui ne traitent pas forcément des mêmes concepts contenus dans la dimension sémantique de son profil. Sinon les recommandations sémantiques doivent être les plus prioritaires puisqu'elles sont plus fidèles aux centres d'intérêts de l'utilisateur

$$u(c,s) = \alpha \cdot u_{coll}(c,s) + \beta \cdot u_{social}(c,s) + \delta \cdot u_{sem}(c,s) \quad [16]$$

Avec $\alpha + \beta + \delta = 1$

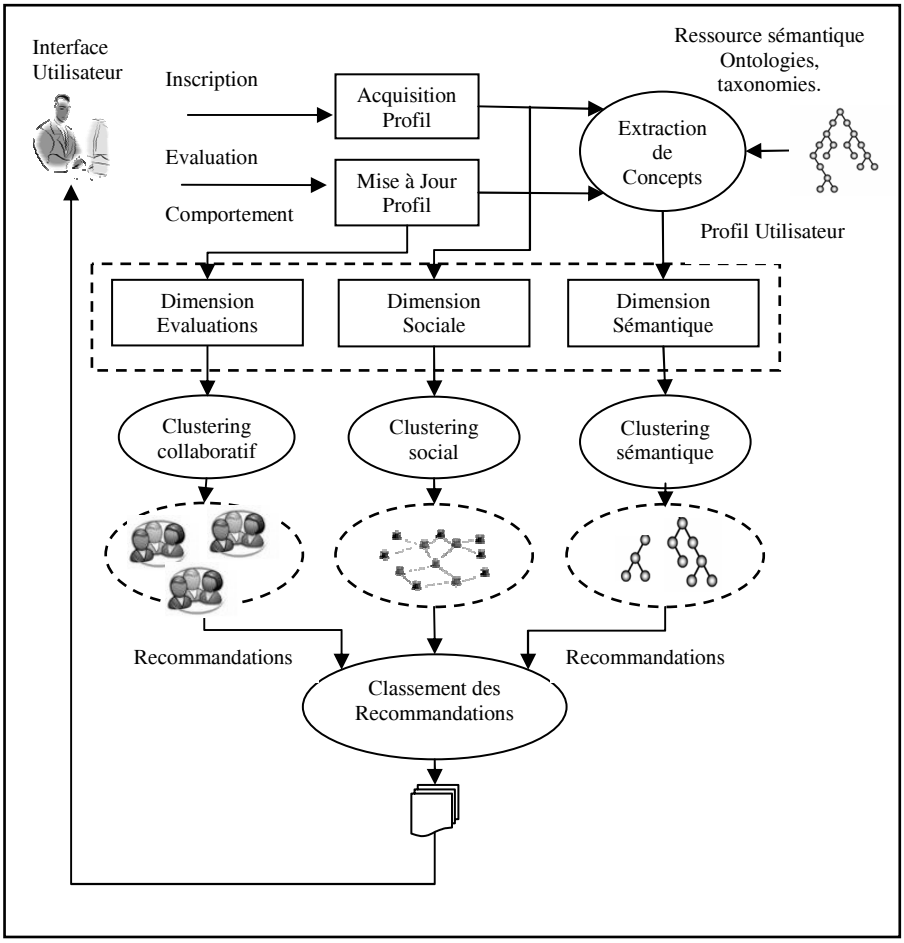

Figure 4. *Un processus de filtrage hybride*

6. Bibliographie

Adomavicius G., Tuzhilin A., "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions", *IEEE Transactions on knowledge and data engineering*, vol. 17, n° 6, 2005.

Baeza-Yates R., Ribeiro-Neto B., *Modern Information Retrieval*, Addison-Wesley, 1999.

Balabanovic M., Shoham Y., "Fab: content-based, collaborative recommendation", *Communications of the ACM*, vol. 40, n° 3, 1997, p. 66-72.

Basu C., Hirsh H., Cohen W., "Recommendation as Classification: Using Social and Content-Based Information in Recommendation", *Proc. of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence AAAI/IAAI*, 1998, p. 714-720.

Baziz M., Boughanem M., Aussenac-Gilles N., Chrisment C., "Semantic cores for representing documents in IR", *Proc. Of the 2005 ACM Symposium on applied computing*, vol. 2, USA, 2005, p. 1011-1017.

Belkin N.J., Croft B., "Information filtering and information retrieval: two sides of the same coin?", *In communications of the ACM*, 1992, p. 29-38.

Billsus D., Pazzani M., "Learning Collaborative Information Filters", *Proc. of the Fifteenth International Conference on Machine Learning*, 1998, p. 46-54.

Breese J.S., Heckerman D. and Kadie C., "Empirical Analysis of Predictive Algorithms for Collaborative Filtering", *Proc. 14th Conf. Uncertainty in Artificial Intelligence*, July 1998.

Chickering D.M., Heckerman D., Meek C., "A Bayesian approach to learning Bayesian networks with local structure", *Proc. of Thirteenth Conference on Uncertainty in Artificial Intelligence*, 1997.

Claypool M., Gokhale A., Mir T., Murnikov P., Netes N., Sartin M., "Combining content-based and collaborative filters in an online newspaper", *Proc. of ACM SIGIR Workshop on Recommender Systems*, 1999.

Claypool M., Brown D., Le P., Waseda M., "Inferring User Interest", *IEEE Internet Computing* 5, 32-39, 2001.

Delgado J. Ishii N., Ura T., "Content-based collaborative information filtering: actively learning to classify and recommend documents", Lecture Notes In Computer Science, vol. 1435, *Proc. of the Second International Workshop on Cooperative Information Agents II*, 1998, p. 206-215.

Dempster A. P., Laird N. M., Rubin D. B., "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society*, Series B, p. 34:1-38, 1977.

Foltz P.W., Dumais S.T., "Personalized Information Delivery: An Analysis of Information Filtering Methods", *Communications of the ACM 35* (12), 51-60, 1992.

Golbeck J., Computing and Applying Trust in Web-based Social Networks, Ph.D. thesis, University of Maryland, College Park, MD, USA, 2005.

Golbeck J., "Generating Predictive Movie Recommendations from Trust in SocialNetworks", *Proc. of the Fourth International Conference on Trust Management*, 2006.

Goldberg D., Nichols D., Oki B. M., Terry D., "Using collaborative filtering to weave an information tapestry", *Communications of the ACM*, vol. 35, n° 12, 1992, p. 61-70.

Groh G., Ehmig C, "Recommandations in taste related domains: Collaborative Filtering vs. Social Filtering", *ACM GROUP'07*, 2007.

Hauger S., Tso K., Schmidt-Thieme L., "Comparison of Recommender System Algorithms focusing on the New-Item and User-Bias Problem", *Proc. of 31th Annual Conference of the Gesellschaft für Klassifikation (GfKl)*, 2007.

Hofmann T., "Latent Semantic Models for Collaborative Filtering", *ACM Trans. Information Systems*, vol. 22, n° 1, 2004, p. 89-115.

Karen H. L., Marinho L.B., Schmidt-Thieme L., "Tagaware Recommender Systems by Fusion of Collaborative Filtering Algorithms", *ACM SAC'08*, 2008.

Khan L., Luo F., "Ontology construction for information selection", *Proc. of the 14th IEEE international conference on tools with artificial intelligence*, USA, 2002, p. 122-12.

Linden G., Brent S.,York J, "Amazon.com recommendations: Item-to-item collaborative filtering", *IEEE internet computing*, vol. 7, n° 1, 2003, p. 76-80.

Lopez M., Accès à l'information par un système de filtrage collaboratif contrôlé, Rapport de thèse de doctorat, université de Joseph Fourier, 2005.

Lops P., Degemmis M., Semeraro G, "Improving social filtering techniques through WordNet-Based user profiles", *User Modeling*, 2007.

Lumineau N., Un tour d'horizon du filtrage collaboratif, Rapport de recherche dans le cadre de l'AS personnalisation de l'information, 2002.

MacQueen J. B., "Some Methods for classification and Analysis of Multivariate Observations", *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, 1, 1967, p. 281-297.

Massa P. and Avesani P., "Trust-aware Collaborative Filtering for Recommender Systems", *On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE*., Berlin, Heidelberg, Springer, 2004, p. 3-17.

Mehta B., Niederee C., Avare S., Degemmis M., Lops P., Semeraro G., "Ontologically-enriched unified user modeling for cross-system personalization", *International conference on user modeling*, vol. 3538, p. 119-123, 2005.

Melville P., Mooney R.J., Nagarajan R., "Content-Boosted Collaborative Filtering for Improved Recommendations", *Proc. 18th Nat'l Conf. Artificial Intelligence*, 2002.

Middleton S.E., Shadbolt N.R., de Roure D.C., "Ontological User Profiling in Recommender Systems", *ACM Trans. Information Systems*, vol. 22, n° 1, 2004, p. 54-88.

Mobasher B., Jin X., Zhou Y., *Semantically enhanced collaborative filtering on the Web*, Book chapter, Web Mining: FromWeb to SemanticWeb, 2004.

Mooney R.J., Bennett P.N., Roy L., "Book Recommending Using Text Categorization with Extracted Information", *Proc. Recommender Systems Papers from 1998 Workshop*, Technical Report, WS-98-08,1998.

Nakamura A., Abe N., "Collaborative filtering using weighted majority prediction algorithms", *In Machine Learning: Proceedings of the Fifteenth International Conference (ICML'98),* Madison, WI, 1998.

Nguyen A., Denos N., Berrut C., "Improving New User Recommendations with Rule-based Induction on Cold User Data", *ACM RecSys'07*, 2007.

Pazzani M., Billsus D., "Learning and Revising User Profiles: The Identification of interesting Web Sites", *Machine Learning*, vol. 27, 1997, p. 313-331.

Pazzani M., "A Framework for Collaborative, Content-Based, and Demographic Filtering", *Artificial Intelligence Rev*, p. 393-408, Dec. 1999.

Peis E., Morales-del-Castillo J. M., Delgado-López J. A., "Semantic Recommender Systems. Analysis of the state of the topic", *Hipertext.net*, n° 6, 2008.

Resnick P., Iacovou N., Suchak M., Bergstrom P., Riedl J, "GroupLens: An open architecture for collaborative filtering of netnews", *Proc. of the 1994 Conference on Computer Supported Collaborative Work*, Furuta, R. and Neuwirth, C., Eds. ACM Press, New York, 1994, p. 175-186.

Salton G., *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley, Reading, MA, 1989.

Sarwar B., Karypis G., Konstan J., Reidl J., "Item-based collaborative filtering recommendation algorithms", *Proc. of the 10th international conference on World Wide Web*, 2001, p. 285-295.

Schein A. I., Popescul R., Ungar L.H., Pennock D., "Methods and metrics for cold-start recommendations", *Proc. of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2002.

Schickel-Zuber V. and Faltings B., "Inferring User's Preferences using Ontologies", *AAAI 2006*, 2006, p. 1413-1418.

Shardanand U., Maes P., "Social Information Filtering: Algorithms For Automating 'Word Of Mouth", *Proc. Conf. Human Factors In Computing Systems*, 1995.

Sinha R., Swearingen K., "Comparing recommendations made by online systems and friends", *DELOS-NSF Workshop on Personalization and Recommender Systems in Digital Libraries*, 2001.

Szomszor M., Cattuto C., Alani H., O'Hara K., Baldassarri A., Loreto V., Servedio V.D.P., "Folksonomies, the Semantic Web, and Movie Recommendation", *Proc. of the ESWC'07*, 2007.

Ungar L.H., Foster D.P., "Clustering Methods for Collaborative Filtering", *Proc. Recommender Systems*, Papers from 1998 Workshop, Technical Report WS-98-08, 1998.

Woerndl W., Eigner R., "Utilizing Physical and Social Context to Improve Recommender Systems", *Workshop on Web Personalization and Recommender Systems (WPRS07)*, USA, 2007.

Yu K., Schwaighofer A., Tresp V., Xu X. and H.-P. Kriegel, "Probabilistic Memory-Based Collaborative Filtering", *IEEE Trans. Knowledge and Data Eng.*, vol. 16, n° 1, Jan. 2004, p. 56-69.

Zhang Y., Callan J., "Maximum Likelihood Estimation for Filtering Thresholds", *Proc. 24th Ann. Int'l ACM SIGIR Conf.*, 2001.

Zhang Y., Callan J., Minka T., "Novelty and Redundancy Detection in Adaptive Filtering", *Proc. 25th Ann. Int'l ACM SIGIR Conf.*, 2002, p. 81-88.

Ziegler C.-N., Schmidt- Thieme L. and Lausen G, "Exploiting semantic product descriptions for recommender systems", *ACM SIGIR Semantic and Information Retrieval Workshop*, 2004.

Ziegler C.-N., Golbeck J., "Investigating interactions of trust and interest similarity", *Decision Support Systems*, vol. 43, n° 2, 2007, p. 460-475.