République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université IBN KHALDOUN- TIARET
Faculté des Sciences de l'Ingénieur

N° d'ordre : | | | | | | | | | |

École Doctorale – STIC

Mémoire

Présenté en vue de l'obtention du diplôme de
Magistère en informatique

Option : Systèmes d'**I**nformation et de **C**onnaissance (SIC)

THÈME

Filtrage collaboratif basé items : Intégration des données sémantiques et optimisation de la prédiction

Présenté par : M^{r} KHARROUBI Sahraoui

Dirigé par : Dr. NOUALI Omar
Maître de recherche CERIST

Jury

Remerciement

Tout d'abord, je tiens à remercier vivement et profondément, Mr NOUALI directeur de mémoire pour avoir encadré ce sujet et pour le temps et l'attention qu'il a bien voulu consacrer au déroulement de ce projet malgré ses occupations professionnelles et ses tâches administratives.

Je remercie vivement les membres de jury pour avoir accepter de juger ce modeste travail.

Enfin, je remercie tous ceux qui de près ou de loin ont bien voulu m'encourager pour que ce travail puisse être achevé.

À

La mémoire de mon père …

Résumé

La masse gigantesque de l'information sur la toile ne cesse d'augmenter, elle est mesurée en péta octet voire en exa octet. Il est nécessaire de développer les techniques de filtrage pour fournir l'information pertinente selon un centre d'intérêt spécifique. Un système de filtrage collaboratif (SFC) génère des recommandations aux utilisateurs par le biais de la similarité et la proximité entre les profils ainsi que par la prise en compte de leurs historiques d'évaluations. A l'inverse de la plupart des SFC qui sont basés sur l'exploitation de la similarité entre utilisateurs, nous adoptons l'approche basée items afin d'améliorer la qualité de recommandation, ce procédé semble flexible et nous a permis d'intégrer d'autres sources d'informations tout en mettant le calcul en mode off-line, puis et pour augmenter les performances et alléger l'inconvénient du manque évaluation, nous avons exploité la couche sémantique des objets telles que les méta-données et les relations sémantiques entre les items, finalement nous avons exploré la technique LSI (Latent Semantic Indexing) afin de réduire la complexité d'algorithme et identifier les items les plus corrélés. Un jeu de test réel de MovieLens á été exploité pour les tests d'expérimentation.

Mots clés : filtrage collaboratif, les systèmes de recommandation, profil utilisateur, sémantique, méta_ données.

Abstract

The huge mass of information on the Web is increasing, it is measured in peta-byte even consider exa-byte. It is necessary to develop filtering techniques to provide information relevant to a specific interest. A collaborative filtering system (CFS) makes recommendations to users via the similarity and proximity between the profiles and taking into account their historical valuations. In contrast to most of the SFC, which are based on the approach-based users, we adopt the approach based items to improve the quality of recommendation, this process seems flexible and allowed us to integrate other sources of information while making the calculation mode OFF-LINE, and then to improve performance and reduce the inconvenience of the lack evaluation, we used the semantic layer objects such as meta-data and semantic relationships between items finally we explored the technique LSI (Latent Semantic Indexing) to reduce the complexity of algorithm and identify the items most corollas. A set of real MovieLens test has been used for experimental tests.

Keywords*:* Collaborative filtering, recommender systems, user profiling, semantic, metadata.

Table des matières

Liste des figures & tables

Partie I

État de l'art

Introduction générale

Les études statistiques et selon Internet World Stats, le nombre mondial d'internautes s'élève à 1,085 milliard en septembre 2006, soit 16,7% de la population mondiale. Le nombre de pages Web accessibles a augmenté de 320 millions en 1997 à plus de 3 milliards en 2002. Le volume d'informations disponible sur Internet ne cessant d'augmenter chaque jour ce qui conduit au problème de ***surcharge*** de l'information. D'une autre part les récents progrès des technologies de l'information de manière générale, des réseaux de communication de manière particulière, ont redonné à l'information de nouveaux contours et d'avantage de valeur selon divers aspects : scientifique, technique, économique, d'usage etc. De surcroît, les progrès techniques de numérisation et de compression de l'information ont encouragé sa production sa circulation et son exploitation. D'autre part être informé étant une nécessité professionnelle et citoyenne, recevoir des informations ayant un certain niveau d'intérêt individuel permet à chacun d'apprendre, d'analyser, de critiquer toute nouvelle source d'information. Ainsi recevant toute nouveauté, l'utilité du filtrage permet donc d'éviter de procéder régulièrement à une recherche d'éventuelles avancées, cela procure à l'utilisateur bien évidemment une économie d'effort mais également une certaine sérénité [BER03].

Il devient donc de plus en plus nécessaire de développer des outils permettant de filtrer cette masse gigantesque d'information, pour cibler au mieux les réponses fournies aux demandeurs, afin qu'elles soient plus proches de leurs besoins et attentes. En effet, la phase de recherche d'information s'appuie particulièrement sur la manière d'accéder aux informations via des requêtes ou par navigation, on assiste aujourd'hui de plus en plus à la profération de services qui ramène des informations à l'utilisateur. Certes, les systèmes de recherche d'information sont des outils qui ont permis d'améliorer sans cesse la qualité des services d'accès à l'information, grâce à la capitalisation des théories issues de nombreux travaux de recherche cependant, en raison de la surabondance de l'information d'une part et de sa large accessibilité à travers le Web, d'autre part, leur mise en œuvre est confrontée à de nouveaux problèmes.

La situation est actuellement paradoxale c'est-à-dire que la masse d'informations est telle que l'accès à une information pertinente adaptée aux besoins d'un utilisateur donné devient à la fois difficile et nécessaire. Le problème n'est pas tant la disponibilité de l'information mais sa pertinence relativement à un contexte d'utilisation spécifique.

C'est pourquoi les travaux s'orientent actuellement vers la révision de la chaîne d'accès à l'information dans la perspective d'intégrer l'utilisateur comme composante du modèle global de recherche et ce, dans le but de lui délivrer une information pertinente adaptée à ses besoins précis, son contexte et ses préférences. Ces travaux s'inscrivent dans le cadre précis de la personnalisation de l'information qui est vue comme l'une des solutions pouvant maintenir le Web comme une ressource viable [GOW03].

Le processus qui permet de sélectionner l'information désirée dans ces flots d'informations s'appelle le filtrage d'information. Un système de filtrage d'information permet à partir d'une source dynamique d'information (Internet, E-mail, News,…) de sélectionner et de présenter les seuls documents intéressant á un utilisateur ayant un centre d'intérêt relativement stable appelé profil. Le filtrage d'information est un processus dual à la recherche d'information [BEL92].

De façon générale, les systèmes de filtrage collaboratif SFC aident un utilisateur à trouver l'information qui l'intéresse à partir des jugements d'autres utilisateurs. Cette approche a pour but d'automatiser les recommandations que peuvent se faire des personnes partageant les mêmes centres d'intérêts.

Problématiques

- Traditionnellement, le SFC compare une représentation d'un utilisateur actif avec les autres voisins semblables pour la décision de recommandation. Ces techniques orientées utilisateur souffrent de quelques limitations critiques telles que l'évolutivité de la mémoire du fait que la phase de formation de voisinage soit exécutée en ligne, ce problème est encore accentué dans le cas du web où le système gère un grands nombre d'usagers.

- Une autre limitation face à ces systèmes, quand le nombre d'items (articles) augmente dans la base, le pourcentage d'évaluation de chaque utilisateur concernant ces items est diminue ce qui baisse le facteur de similarité dans le voisinage, ceci à son tour dégrade la qualité de recommandation.

- En outre, le SFC peut ne jamais produire de recommandations pour de nouveaux items qui n'ont pas été encore évalués par (nombre suffisant) d'autres utilisateurs, ce phénomène connu sous le nom de "démarrage à froid".

Contributions

Afin d'améliorer l'efficacité et augmenter les performances des SFC nous avons choisi la démarche suivante :

Premièrement l'adoption de l'approche basée items pour l'algorithme de filtrage collaboratif à l'inverse des systèmes classiques qui sont basés sur l'approche user (mode on-line) pour le calcul de la similarité.
Cette approche suggérée est caractérisée par :

- Le calcul de la similarité entre items se fait en mode off-line (temps régulier, mode batch,…) ce qui augmente les performances du système.
- L'espace des items est relativement faible par rapport á l'espace des utilisateurs ce qui raccourci la similarité entre ces items.

Deuxièmement, dans l'absence de données d'évaluation, nous avons proposé l'intégration de l'information sémantique partagée entre les items (méta données, sens de relations entre items …), cette approche permet de :

- Réduire le problème de manque de données dans la base et par conséquent l'amélioration de la qualité de prédiction.
- Pour les nouveaux items, le système fait une affectation dans le voisinage le plus similaire selon l'information sémantique commune entre ces items.

En fin, dans le but de réduire la complexité de calcul, nous avons optimisé notre approche par l'exploitation de la technique LSI (Latent Semantic Indexing) pour réduire l'espace initial de traitement tout en gardant les items les plus corrélés.

L'organisation de mémoire

Ce mémoire est organisé en quatre chapitres :

Le chapitre I, ***Filtrage d'information,*** rappel les principes de base des systèmes de filtrage d'information, ainsi que les différentes approches existantes actuellement. Le propos de ce chapitre est de présenter les éléments généraux composant les systèmes de filtrage collaboratif,

Le chapitre II, ***web sémantique et filtrage d'information***, nous avons présenté une vision globale du web de demain, où nous pouvons effectuer plusieurs tâches de manière informatisée et automatisée grâce à l'explicitation de la sémantique des ressources et des connaissances, ensuite nous avons exploré les éléments de base de cette infrastructure pour améliorer l'efficacité et les performances des SFC objet central de notre travail.

Le chapitre III, ***Objectifs et proposition,*** consacré à notre approche qui est basée essentiellement sur trois points de vue, en premier lieu l'adoption de l'approche basée items versus l'approche basée utilisateur utilisée dans les systèmes traditionnels, puis nous avons intégré les données sémantiques pour progresser la fonction de prédiction, finalement, nous avons introduire une technique d'optimisation LSI dans l'algorithme de génération de recommandations.

Le chapitre IV, ***Expériences et résultats,*** on à présenté les mesures d'évaluations standards pour les SFC ensuite nous avons évalué notre solution sur un jeu de test réel de MovieLens.

Annexe A, un glossaire des commandes MATLAB et un listing des algorithmes utilisés.

Annexe B, l'interface graphique réalisée en java pour la visualisation des différents résultats de notre approche proposée.

CHAPITRE I
FILTRAGE D'INFORMATION

L'objectif de ce chapitre est de présenter un état de l'art consacré aux systèmes de filtrage (SF), puis nous décrivons les techniques existantes les plus répandues, Enfin, nous focalisons sur les systèmes de filtrage collaboratif (SFC) en discutant les processus et les facteurs principaux ainsi que leurs limitations.

1. Enjeux et définitions

1.1 Facteurs dominants

Parmi les facteurs dominants qui constituent des enjeux actuels dans le domaine du filtrage d'information, on retient : le volume, l'hétérogénéité et disparité des informations [LYN05].

Le volume

Ainsi, le volume d'informations ne se mesure plus actuellement en giga-octets mais en téraoctets voire en péta octets et exa-octets. Cette croissance des volumes de stockage engendre le problème d'allongement des délais de réponses, l'augmentation des coûts d'indexation ainsi que la diminution de la précision de la recherche.

Hétérogénéité

Porte sur divers aspects : la diversité des langues et le type d'information (texte, image, vidéo), structure, les organisations, etc. Le Web sémantique est une nouvelle couche qui permet d'annoter les ressources du web pour augmenter le facteur de précision.

Disparité

Une caractéristique qui traduit l'occurrence disséminée de l'information dans de larges collections de documents, compte tenu du volume important d'informations disponibles, les utilisateurs sont vite submergés par le nombre considérable de liens proposés, ce qui engendre les phénomènes fort connus de désorientation de l'utilisateur et de surcharge informationnelle.

1.2 Définition de l'information

D'un point de vue scientifique, l'information apparaît comme un sujet vague et incohérent [MEK94].

De ce fait, le mot « information » a des définitions multiples et ambiguës. D'après "Larousse", elle paraît tout à fait significative. Elle se décompose en plusieurs sous définitions selon les critères suivants :

- Action : «*L'information est l'action d'informer, de se mettre au courant d'événements*».
- Etat : «*L'information est une nouvelle, un renseignement que l'on communique ou que l'on obtient*».
- Connaissance : «*L'information est un ensemble de connaissances acquises sur quelqu'un ou sur quelque chose*».
- Contenu : «*L'information est le contenu proprement dit des messages transmis*».
- Contenant : «*L'information est un signal par lequel un système donne connaissance de sa position à un autre* ».

La première grande contribution à la théorie de l'information a été faite par Claude Shannon en 1948. Il propose de considérer l'information comme une quantité physique, au même titre qu'une masse ou que de l'énergie, transmise à l'aide d'un canal, et procurant une codification de l'état du monde parmi plusieurs états possibles. Étudier l'information c'est étudier l'information proprement dite (la quantité d'information, l'entropie d'une source d'information), les propriétés des canaux (transmission équivoque, bruit, capacité…) et enfin, les relations qui existent entre l'information à transmettre et le canal employé, en vue d'une utilisation optimale de celui-ci [MIC99].

D'autres définitions de l'information apparaissaient dans les années 50/60 dans le champ des sciences humaines par des philosophes, des psychologues, des cybernéticiens et des sémioticiens. Selon les premiers, l'information se définit exclusivement au niveau de l'esprit à travers l'accumulation de messages. Pour les cybernéticiens, tous les organismes comme les plantes, les animaux, les humains mais aussi les machines ou les ordinateurs, sont des

émetteurs ou récepteurs de signes, alors que pour les sémioticiens, l'information est de nature sémantique : il existe une relation claire entre le sens de l'enregistrement et l'information qu'il produit.
Dans la recherche d'information, nous rejoignons la définition de Coadic : « *L'information est une connaissance inscrite (enregistrée) sous forme écrite (imprimée ou numérisée), orale ou audiovisuelle.*
L'information comporte un élément de sens. C'est une signification transmise à un être conscient par le moyen d'un message inscrit sur un support spatiotemporel : imprimé, signal électrique, onde sonore, etc. Cette inscription est faite grâce à un système de signe (le langage), le signe étant un élément du langage qui associe un signifiant à un signifié : signe alphabétique, signe de ponctuation. Le but de l'information reste l'appréhension de sens ou d'êtres dans leur signification, jusqu'à le reste de la connaissance, la transmission du support, de la structure en étant le moyen. » [LEC94].

1.3 Recherche d'information RI

La Recherche d'Information **(RI)** est un domaine de l'informatique qui s'intéresse à l'organisation, au stockage et à la sélection d'informations répondant aux besoins des utilisateurs [SAL70], [SAL84]. Ce domaine manipule différents concepts : la requête, le besoin en information, les documents, la pertinence, etc. Deux types de systèmes peuvent donner accès à l'information : la collecte active à travers les systèmes de recherche d'information **(SRI)** et la collecte passive à travers les systèmes de filtrage d'information **(SFI).**

• Le processus de RI

Le processus de recherche d'information est le processus qui permet de mettre en relation l'ensemble des informations disponibles d'une part et les besoins de l'utilisateur d'une autre part.
L'expression de ces besoins se fait par le biais de requêtes. Ces requêtes envoyées au système de recherche d'information afin qu'il extrait les documents pertinents répondant au besoin de l'utilisateur. Cette notion de pertinence est fortement subjective, car elle dépend de l'utilisateur, donc très difficile à automatiser.
Le processus de recherche d'information comprend plusieurs concepts :

- *La collection de documents (corpus):* un corpus de documents est un ensemble de granules documentaires qui peuvent être des documents entiers ou bien des parties de documents qui est considéré comme unité d'information dans la RI classique.

- *Le besoin en information:* toute interrogation, vœux, intérêt d'un utilisateur est exprimé en langage naturel booléen ou graphique via une requête (ensemble de mots clés) acheminée par un SRI afin de rendre l'information à l'utilisateur.

- *La fonction d'indexation :* l'indexation est le processus permettant de créer une représentation des documents et des requêtes facilement manipulable par un système de recherche d'information. Elle consiste à analyser les documents afin d'extraire un ensemble de mots clés servant comme descripteurs des documents. Elle permet la représentation des documents sous une forme réduite et succincte [SID02] et évite au système de recherche d'informations l'exploration de tout le contenu du document à chaque nouvelle interrogation [MKA04]. Il existe trois types d'indexation :

a) Indexation manuelle : l'extraction et le choix des descripteurs s'effectuent par un documentaliste ou un spécialiste du domaine.

b) Indexation automatique : l'extraction et le choix des descripteurs s'effectuent d'une façon totalement automatisée.

Remarque : avec l'expansion du web et la masse d'information et le besoin de la rapidité, l'indexation automatique s'avère nécessaire. Cependant l'indexation automatique n'est pertinente qu'à 60% contre 95% pour l'indexation humaine qui exige la compétence des analystes, cette différence réside dans l'ambiguïté de la langue et le domaine d'application [ELH97].

c) Indexation semi-automatique l'extraction des descripteurs s'effectue par le système et le choix des descripteurs est laissé au spécialiste [MAT07].

- *La fonction d'appariement requête-document:* cette fonction est définie afin de mesurer la pertinence d'un document vis-à-vis d'une requête. Elle est vue comme une probabilité ou une similarité vectorielle, notée RSV (Q, d) *(Retrieval Status Value)*, où Q représente la requête et d représente un document. La fonction de similarité permet d'ordonner les documents renvoyés à l'utilisateur par ordre de pertinence.

-*La fonction de modification de requête* : L'expression du besoin en information d'un utilisateur sous forme de requête est souvent une chose difficile. Par conséquent, les documents trouvés par la requête initiale peuvent ne pas accomplir le besoin en information de l'utilisateur. C'est pour cette raison que le système de recherche d'information fait appel à la fonction de modification de requêtes afin de corriger le chemin de la recherche. Une fois que le système présente un premier ensemble de documents, ils peuvent facilement différencier entre les documents qui contiennent de l'information utile et ceux qui ne la contiennent pas. C'est ce qu'on appelle communément la réinjection de pertinence (*relevance feedback*), l'une des techniques de modification de requêtes basée sur la pondération sur les termes retrouvés dans les documents pertinents, et l'opération se répète jusqu'à la satisfaction de l'utilisateur.

- *La fonction d'appariement requête-document:* cette fonction est définie afin de mesurer la pertinence d'un document vis-à-vis d'une requête. Elle est vue comme une probabilité ou une similarité vectorielle, notée RSV (Q, d) *(Retrieval Status Value)*, où Q représente la requête et d représente un document. La fonction de similarité permet d'ordonner les documents renvoyés à l'utilisateur par ordre de pertinence.

-*La fonction de modification de requête* : L'expression du besoin en information d'un utilisateur sous forme de requête est souvent une chose difficile. Par conséquent, les documents trouvés par la requête initiale peuvent ne pas accomplir le besoin en information de l'utilisateur. C'est pour cette raison que le système de recherche d'information fait appel à la fonction de modification de requêtes afin de corriger le chemin de la recherche. Une fois que le système présente un premier ensemble de documents, ils peuvent facilement différencier entre les documents qui contiennent de l'information utile et ceux qui ne la contiennent pas. C'est ce qu'on appelle communément la réinjection de pertinence (*relevance feedback*), l'une des techniques de modification de requêtes basée sur la pondération sur les termes retrouvés dans les documents pertinents, et l'opération se répète jusqu'à la satisfaction de l'utilisateur.

La reformulation de requêtes peut s'effectuer selon deux stratégies :
L'extension de la requête avec de nouveaux termes, et la répondération des termes de la requête initiale.
La modification de la requête peut être manuelle, automatique ou bien semi-automatique [MAT07].

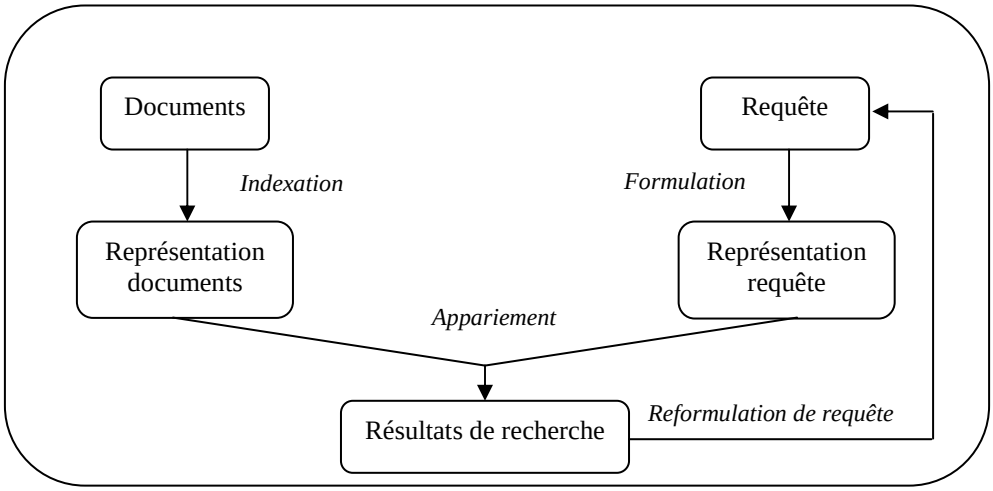


Fig I.1- Processus de recherche d'information [RIJ79]

● Les modèles de RI

Un modèle de RI a pour rôle de fournir une formalisation du processus de recherche d'information. Il doit accomplir plusieurs rôles dont le plus important est de fournir un cadre théorique pour la modélisation de la mesure de pertinence [SAL83]. On distingue trois principaux modèles : ensemblistes, algébriques et probabilistes.

- Les modèles ensemblistes

Les modèles ensemblistes reposent sur la théorie des ensembles. Dans ces modèles, les termes de la requête sont séparés par des opérateurs logiques : conjonction (ET), disjonction (OU) et négation (NON). Ces opérateurs permettent d'effectuer des opérations d'union « OU », d'intersection « ET » et de différence « NON » entre les ensembles de résultats associés à chaque terme.

- Les modèles algébriques

Les modèles algébriques se basent sur la théorie algébrique. Dans ces modèles, la pertinence d'un document vis-à-vis d'une requête est définie par des mesures de distance (ou similarité) dans un espace vectoriel.

- Les modèles probabilistes

Les modèles probabilistes se basent sur la théorie des probabilités. Pour ces modèles, la pertinence d'un document vis-à-vis d'une requête est vue comme une probabilité de pertinence document/requête.

A titre d'indication il existe d'autres modèles en citant le modèle booléen étendu, le modèle vectoriel généralisé et le modèle de langage [GAL05].

1.4 Filtrage d'information FI

Parmi les définitions qui sont données**,** le FI 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 [BEL92]. Le filtrage peut être vu comme la sélection d'informations pertinentes sur un flux entrant.

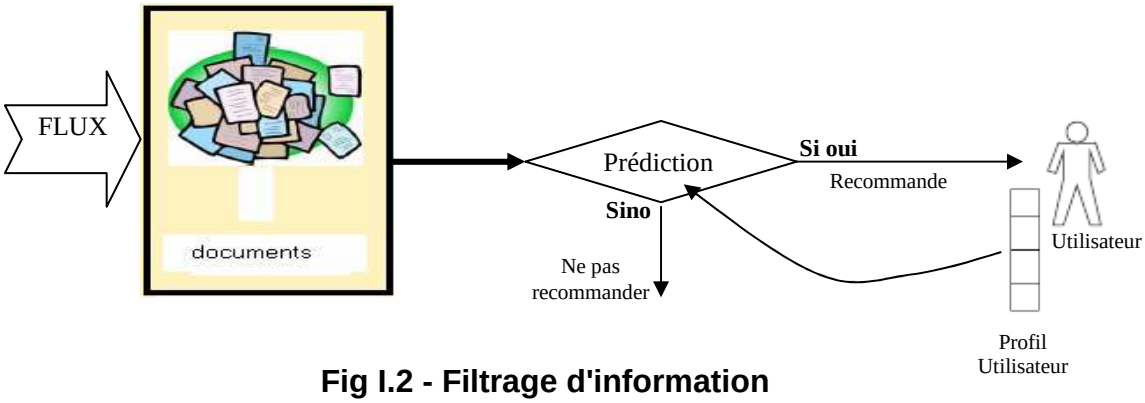


Fig I.2 - Filtrage d'information

Pour sélectionner les informations pertinentes de « Flux entrant », le système fait une « prédiction » quant à l'intérêt que présente l'information pour l'utilisateur. Cette prédiction s'appuie sue le « profil » de cet utilisateur et aboutit à une prise de décision « recommander » ou « ne pas recommander » l'information.
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 FI est réalisé à partir d'un volume important d'informations disponibles dynamiquement [TER93] via le « flux entrant ». Ces informations proviennent éventuellement de sources différentes.
Selon Tmar [TMA02], un système de filtrage d'information est défini par son modèle de représentation des profils utilisateurs, son modèle de représentation des documents et sa fonction de prédiction sur la pertinence des documents reçus. Tout système de filtrage est également régi par le type d'utilisateurs visé ainsi que le type de documents manipulés.

2. Filtrage cognitif

2.1 Description générale

Le filtrage cognitif, nommé aussi basé sur le contenu (Content-based Filtering) [LAN95], est une évolution générale des études sur le filtrage d'information, s'appuie sur le contenu des documents (thèmes abordés) pour les comparer à un profil lui-même constitué de thèmes. Chaque utilisateur du système possède alors un profil qui décrit ses propres centres d'intérêt. Par exemple, le profil peut contenir une liste des thèmes que l'utilisateur aime bien ou qu'il n'aime pas [SAL88]. Lors de l'arrivée d'un nouveau document, le système compare la représentation du document avec le profil pour prédire la satisfaction de l'utilisateur sur ce document.

Le système fait des prédictions de l'opinion qu'un utilisateur aura d'un document donné est calculée en rapprochant les thèmes énoncés par l'utilisateur comme constituant son profil, et les thèmes extraits des documents par un processus d'indexation [MAL95].

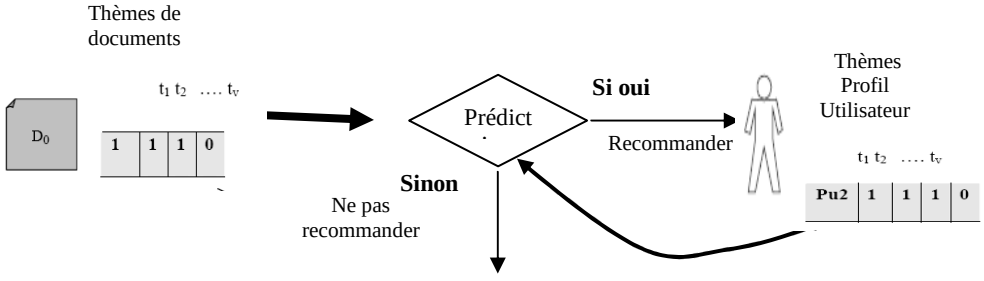


Fig I.3 - Filtrage cognitif

2.2 Limites du filtrage cognitif

Les inconvénients majeurs de ce type de filtrage sont :

• La difficulté d'indexation de documents multimédias (son, image, vidéo) du fait que le filtrage par le contenu se base sur un profil qui exprime sur un plan thématique.

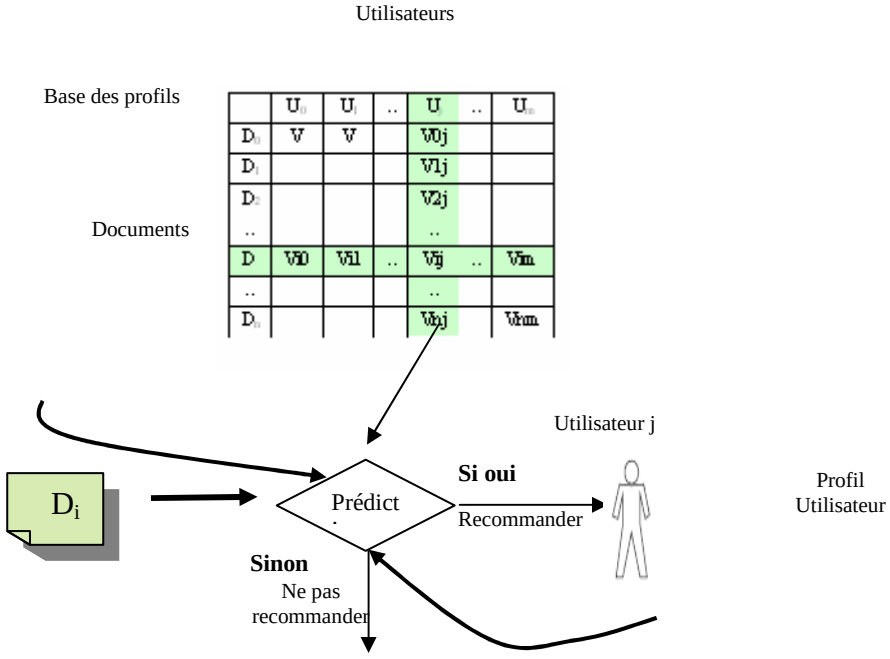
• La restriction sur le critère thématique et exclusion d'autre facteurs de pertinence car l'utilisateur ne bénéficiant pas des jugements de qualité que d'autres utilisateurs ont pu faire sur les documents qu'il reçoit, c'est lui-même qui devra procéder à l'écrémage des documents reçus, écrémage qui fait intervenir d'autres critères que celui de la thématique.

• L'effet de l'entonnoir, car le profil évolue naturellement par restriction progressive sur les thèmes recherchés. Ainsi, l'utilisateur ne reçoit que les recommandations relatives aux thèmes présentés dans son profil, une fois devenu stable. Par conséquent, il ne peut pas découvrir de nouveaux domaines potentiellement intéressants pour lui [COC06], et le système exclut tout genre de similitude.

3. Filtrage collaboratif

3.1 Description générale

Le filtrage collaboratif (Collaborative Filtering) ou social, a pour principe d'exploiter les évaluations que des utilisateurs ont faites de certains documents, afin de recommander ces mêmes documents à d'autres utilisateurs, et sans qu'il soit nécessaire d'analyser le contenu des documents [BHK98].



FigI.4 – Filtrage collaboratif

Pour produire la prédiction d'intérêt qu'un document va susciter chez un utilisateur, le système exploite la « base de profil ». Cette base est une source d'information collaborative qui contient l'intégralité des évaluations faites par les utilisateurs.

• Mode de fonctionnement

Le système calcule la prédiction de l'intérêt que le document d_i va susciter chez l'utilisateur u_j puis l'exploitation des évaluations existantes concernant le document d_i ainsi que le profil de l'utilisateur u_i. La base de profils permet par ailleurs, de connaître l'ensemble des utilisateurs proches d'un utilisateur u_j, et donc de nuancer l'impact des évaluations existantes de d_i selon que les auteurs de ces évaluations sont proches ou non de u_j. Ainsi, la valeur prédite sera d'autant meilleure que les auteurs de « bonnes » évaluations sont proches de u_j.

On note bien l'effet de la proximité entre les utilisateurs du système dans le calcul de la prédiction.
Le profil de l'utilisateur est mis à jour au cours du temps à partir des nouvelles évaluations que l'utilisateur réalise.

Dans le cadre de filtrage collaboratif, les documents ne sont plus qualifiés par leurs contenus (l'approche précédente) mais par les « évaluations » que les utilisateurs en ont faites sur une certaine échelle de valeurs.

Le noyau d'un système de filtrage collaboratif est constitue de trois fonctionnalités : calcul de la proximité entre utilisateurs, calcul de la prédiction i.e la décision de la recommandation (basée sur le profil utilisateur et la proximité) et la mise à jour du profil utilisateur. En outre d'une interface d'édition d'évaluations et visualisation des documents et communautés.

• La prédiction dans le filtrage collaboratif

La prédiction de l'opinion de l'utilisateur sur un document se calcule par rapprochement entre les évaluations passées de l'utilisateur et des autres utilisateurs de la communauté sur les mêmes documents. Le filtrage collaboratif tient compte de la proximité entre les utilisateurs, afin de ne recommander un document qu'aux utilisateurs proches de celui qui l'a apprécié [LEE05].

Remarque : Breese *et al.* [BRE98] propose une classification intéressante des techniques de filtrage collaboratif : les algorithmes basés « mémoire », et les algorithmes basés « modèle ».

3.2 Algorithmes de filtrage collaboratif

3.2.1 Algorithmes basés mémoire

Les algorithmes basés mémoire utilisent l'ensemble de la base de données des évaluations des utilisateurs pour faire les prédictions, les évaluations de l'utilisateur actif sont prédits à partir d'informations partielles concernant cet utilisateur et un ensemble de poids calculés à partir de la base de données des évaluations des utilisateurs.
Si Ii est l'ensemble des items évalués par l'utilisateur i, alors l'évaluation moyenne pour l'utilisateur i peut être définie comme :

$$v_i = \frac{1}{|I_i|} \sum_{j \in I_i} v_{ij}$$

L'évaluation prédite sur l'item j pour l'utilisateur actif a est une somme pondérée des évaluations des autres utilisateurs :

$$P_{a,j} = \overline{v_a} + k\sum_{i=1}^{n} w(a,i)\left(v_{ij} - \overline{v_i}\right)$$

Où n est le nombre d'utilisateurs dans la base de données qui ont un poids non nul, et
k est un facteur de normalisation tel que la somme des valeurs absolues des poids fait 1. Le poids w(a,i) est déterminé de façon variable, selon l'algorithme.

Pour l'algorithme basé sur la corrélation, le poids est calculé comme la corrélation entre les utilisateurs a et i, comme suit :

$$w(a,i) = \frac{\sum_{j}(v_{a,j} - \overline{v_a})(v_{i,j} - \overline{v_i})}{\sqrt{\sum_{j}(v_{a,j} - \overline{v_a})^2 \sum_{j}(v_{i,j} - \overline{vi})^2}}$$

Où les sommes sur les *j* concernent les items pour lesquels à la fois *i* et *a* ont donné des évaluations.

Ces algorithmes ont l'avantage d'être simples à mettre en œuvre, les prédictions sont de bonne qualité car sont recalculées à chaque fois, mais le problème est la forte complexité combinatoire $o(m^2n)$ avec (n : utilisateurs, m : ressources), qui les empêche d'être utilisées dans des environnements de production [AMO08].

3.2.2 Algorithmes basés « modèle »

L'idée de base des algorithmes basés « modèle » est d'employer la base de données des évaluations des utilisateurs pour estimer ou apprendre un modèle qui servira pour le calcul des prédictions. Du point de vue probabiliste, la tâche de prédiction d'une évaluation peut être vue comme le calcul de la valeur espérée d'une évaluation, étant donné ce que l'on sait d'un utilisateur. Supposons que les évaluations se fassent sur une échelle d'entiers de 0 à m.
Alors la valeur prédite sera :

$$P_{a,j} = E(v_{a,j}) = \sum_{i=0}^{m} P_r(v_{a,j} = i \mid v_{a,k} \in I_a)i$$

où la probabilité exprimée est celle dont l'utilisateur actif fera l'évaluation particulière i pour l'item j compte tenu des évaluations observées auparavant.

L'algorithme réduit la complexité et simple pour la mise en œuvre, mais l'évolution de l'apprentissage est lente ainsi que la criticité du choix des classes initiales.
Afin de limiter les inconvénients de ces approches, beaucoup de systèmes adoptent l'hybridation efficace de ces techniques.

3.2.3 Algorithmes hybrides

Une approche graphique [AGG99] repose sur la construction d'un graphe orienté où les utilisateurs sont représentés par les noeuds et l'influence entre les utilisateurs est représentée par les arcs. Cette notion d'influence se décline sous la forme de deux contraintes.
La première, appelée «Horting», donne une information sur le nombre d'évaluations communes entre deux utilisateurs. D'où la définition :

i *horts* k si et seulement si card($R_i \cap R_k$) = min(F.card(R_i) , G)
où F=1 et G sont des seuils

La seconde, dite «prédictabilité», ajoute à la notion de «horting» une information sur le degré de ressemblance entre deux utilisateurs en se basant sur la distance de Manhattan qui les séparent. D'où la définition :

"k *predicts* i" si et seulement si "i *horts* k" et s'il existe une transformation linéaire T telle que DT(i,k) soit inférieur à un seuil prédéfini.

où DT(i,k) représente la distance de Manhattan entre l'utilisateur i et l'utilisateur k à une transformation linéaire près. En d'autre terme, on a :

$$D_T(i,k)=\frac{\sum_{i\in R_i\cap R_k}\left|eval(i,l)-T(eval(k,l))\right|}{card(R_i\cap R_k)}$$

- eval(i,l) est l'évaluation de la ressource l par l'utilisateur i
- T est une transformation linéaire qui à une évaluation définie dans l'ensemble {1,…,v}.
- Ri est l'ensemble des ressources évaluées par l'utilisateur i

Le principe général de cette approche est que la prédiction peut être calculée à partir d'un certain nombre d'arcs qui satisfont les conditions : 'Horting > Seuil1', 'Prédictibilité < Seuil2'
Ceci revient à considérer la transitivité dans la relation de transitive entre utilisateurs. Si un utilisateur 'x' est similaire à 'y', est l'utilisateur 'y' est similaire à 'z' alors la probabilité pour que 'x' et 'y' soient similaires est très grande [AMO08].

3.3 Apport du filtrage collaboratif

Le filtrage collaboratif permet de contourner les limitations du filtrage par le contenu :
• Résout le problème de l'indexation des documents multimédia, du fait que le filtrage est basé sur les évaluations des utilisateurs et non sur le contenu des documents.

• L'émergence d'autres facteurs de pertinence tels que la qualité intellectuelle ou informationnelle, la précision des faits et la fiabilité de la source d'information suivant le facteur de collaboration et contribution entre utilisateurs.

• La crédibilité de l'appréciation humaine par rapport au traitement automatique (difficulté de la langue, synonymie, polysémie) résoudre le problème de l'effet entonnoir.

• Un autre avantage du filtrage collaboratif est que les jugements de valeur des utilisateurs intègrent non seulement la dimension thématique mais aussi d'autres facteurs relatifs à la qualité des documents tels que la diversité, la nouveauté, l'adéquation du public visé, etc. [COC06].

3.4 Limites du filtrage collaboratif

De nombreux systèmes de recommandation s'appuient partiellement ou totalement sur le filtrage collaboratif [BUK02], en raison des avantages importants ci-dessus. On constate néanmoins certains inconvénients de cette technique, incluant le démarrage à froid, la masse critique et le rapport coût-bénéfice.

- Démarrage à froid : Ce phénomène se produit en début d'utilisation du système, dans des situations critiques où le système manque de données pour procéder à un filtrage personnalisé de bonne qualité. En général, la communauté d'un utilisateur évolue au cours du temps grâce aux évaluations produites par l'utilisateur lui-même. Lorsqu'il s'inscrit pour utiliser le système, sa communauté est encore inconnue, ce qui conduit à l'impossibilité de fournir des recommandations pertinentes [COC06];

- Masse critique : Afin de former de meilleures communautés, le système exige un nombre suffisant d'évaluations en commun entre les utilisateurs pour les comparer entre eux. Par exemple, on ne peut pas conclure que deux personnes sont dans une même communauté si elles n'ont qu'une seule évaluation en commun. Et pourtant, vu la taille énorme de l'ensemble des documents, achats, etc. dans les systèmes, le nombre des évaluations en commun entre utilisateurs risque d'être faible [COC06];

- Rapport coût-bénéfice : L'utilisateur évalue l'intérêt du système par rapport à ce qu'il pourrait en tirer en effet lorsqu'il ressent un rapport coût/bénéfice déficitaire ou perçoit une constante baisse dans la pertinence des documents et des recommandations, il peut se décourager ce qui le conduit inévitablement à l'abandon du système [GAL05].

- Les utilisateurs qui sont des centres d'intérêts singuliers, risquent de ne pas recevoir des propositions et des recommandations par le système (peu de similarité entre profils), et par conséquent ils se retrouveront inévitablement isolés des autres utilisateurs du système.

- Le système dépend fortement de son utilisation et du comportement de l'ensemble des utilisateurs, un utilisateur qui ne contribue pas à l'évaluation des documents reçus voit la qualité des réponses et la pertinence des recommandations baisser non seulement pour lui, mais le fait également baisser pour les autres utilisateurs de son voisinage [GAL05].

3.5 Approche collaborative versus approche cognitif

Les deux approches se complètent et le tableau ci – dessous explique les caractéristiques de chaque une des approches.

	Filtrage basé sur le contenu sémantique	**Filtrage collaboratif**
Amorçage (démarrage de l'exploitation du système)	Le filtrage peut commencer après l'établissement du profil	Exige une base de données substantielle et plusieurs évaluations de l'utilisateur avant d'être utilisable
Qualité de l'information (lisibilité, fiabilité, nouveauté, etc.)	La qualité de l'information n'est pas connue	La qualité de l'information est connue *via* des évaluations d'utilisateurs
Contexte de l'information (domaine d'intérêt)	L'identification du domaine se fait généralement par la co-occurrence des termes dans chaque document	L'identification du domaine se fait par la différence des domaines d'intérêt des utilisateurs

Effet « entonnoir »	Le système ne suggère que des documents dont le thème a déjà été évoqué explicitement	Le système peut suggérer des documents sans rapport explicite avec les thèmes déjà évoqués

TabI.1- Comparaison entre filtrage collaboratif et filtrage par le contenu [BER03]

4. Autres formes de filtrage

Filtrage hybride

Constatant les avantages et inconvénients de chacune des deux approches ci-dessus, on comprend que de nombreux systèmes reposent sur leur combinaison, ce qui en fait des systèmes de filtrage dits « hybrides », Plus généralement, les systèmes hybrides gèrent des profils d'utilisateurs orientés contenu, et la comparaison entre ces profils donne lieu à la formation de communautés d'utilisateurs permettant le filtrage collaboratif [COC06].

Filtrage actif

L'atout de ce type de filtrage est la contribution à réduire le démarrage à froid par la possibilité offerte aux utilisateurs de la communauté de se recommander mutuellement des documents. Lorsqu'un utilisateur trouve des documents plus ou moins intéressants pour certains autres utilisateurs qu'il connaît, il peut les leur recommander [DEN04].

Filtrage adaptatif

La plupart des systèmes de filtrage d'information sont basés sur des modèles de recherche d'information. Ainsi, les documents et les profils sont représentés par des listes de mots pondérés. L'appariement document-profil consiste à mesurer une similarité. La décision quant à l'acceptation ou le rejet d'un document est assurée par une fonction de décision souvent de type seuil. Si le score est supérieur au seuil le document est accepté sinon il est rejeté. Or, en l'absence de base de référence, la détermination de ce seuil et les pondérations adéquates associées aux profils et aux documents sont les problèmes majeurs rencontrés dans ce domaine.

A l'initialisation du processus de filtrage, on ne dispose d'aucune connaissance sur les documents à filtrer pour pouvoir construire une fonction de décision, ni pour bien pondérer les profils et les documents entrants. De plus, l'adaptation des profils aux différents flots pose un problème fondamental lié à l'incomplétude permanente des informations permettant de décrire exhaustivement les profils. Parmi les solutions proposées: une solution synchrone ou adaptative, les différents facteurs du système de filtrage sont déduits à partir des documents filtrés cumulés dans le temps, et les solution asynchrone ou différée (filtrage différé), les facteurs sont déduits à partir de collections de documents existantes.

Aucune information autre que le profil initial n'est connue au démarrage du processus de filtrage. Les statistiques des documents et des profils sont actualisées au fur et à mesure que le système reçoit des documents. L'adaptation des profils et le seuillage se font d'une manière adaptative et incrémentale à chaque réception d'un document pertinent.

L'adaptation du profil est basée sur le principe de renforcement. Chaque fois qu'un document est sélectionné et jugé pertinent pour un profil donné, le système doit adapter le profil de sorte à modifier sa représentation. De plus, la méthode d'adaptation du seuil basée sur la distribution des scores d'un échantillon de documents pertinents et non pertinents sélectionnés [TMA01].

5. Profil utilisateur

La pertinence de l'information se définit par un ensemble de critères et de préférences personnalisables spécifiques à chaque utilisateur ou communauté d'utilisateurs représentées par des couples *(attribut, valeur)*. Les données décrivant les utilisateurs sont souvent regroupées sous forme de profils**.** Le contenu du profil d'un utilisateur varie selon les approches et les applications. Les approches existantes répondent partiellement aux questions liées à la personnalisation, mais il manque un modèle donnant une vision globale sur tous les aspects de la prise en compte des préférences des utilisateurs [BOZ04].

La représentation de l'utilisateur à travers la notion de profil permet de mieux comprendre ses mécanismes cognitifs, notamment ceux permettant de percevoir le concept subjectif de la pertinence et au-delà, cibler ses besoins spécifiques dans le but d'améliorer la pertinence de l'information. [DAN86] définit deux classes de modèles de profils utilisateurs :

. Les modèles quantitatifs et empiriques : leur but est de modéliser le comportement externe de l'utilisateur,

. Les modèles analytiques et cognitifs : leur but est de comprendre le comportement interne de l'utilisateur : connaissance, raisonnement, etc.

Ces deux aspects sont généralement combinés pour représenter, construire et faire évoluer les profils.

Il est important de noter que dans le principe du filtrage purement collaboratif le profil de l'historique des évaluations ne comporte pas d'informations sur le contenu des documents évalués mais seul l'identificateur du document est conservé dans le profil. Par contre, dans un système de recommandation hybride combinant le filtrage collaboratif avec d'autres techniques. Chaque utilisateur possède un profil « multidimensionnel » qui comprend, outre l'historique des évaluations, d'autres données telles que les informations personnelles, les centres d'intérêt, etc.

5.1 Modélisation du Profil

L'introduction de la dimension utilisateur dans un processus d'accès à l'information mérite voire nécessite des réflexions sur la modélisation de l'entité *utilisateur* [GOW03]. Dans ce sens, les questions fondamentales posées par la conception de tels systèmes sont de type Quoi, Comment et Quand :

Quoi ?

. Quelles propriétés décrivent un utilisateur ?
. Quelle représentation ou quel modèle de l'utilisateur ?
. Quel contexte d'utilisation ?

Comment ?

. Comment construire le modèle de l'utilisateur ?
. Comment découvrir son intention courante ?
. Comment exploiter le modèle utilisateur lors du processus d'interaction ?
. Comment exploiter le retour de pertinence *" feedback"*

Quand ?

. Quand faut il mettre à jour le modèle de l'utilisateur ?

Indépendamment du domaine d'application, tout système mettant en œuvre des méthodes de modélisation de l'utilisateur inclut en partie les paquets d'informations suivants [LYN05] :

. Des informations personnelles associées à l'utilisateur telles que l'âge, le pays, la langue,

. Les préférences : peuvent être de différents niveaux telles que préférences de forme (style de la page, longueur d'un document) et préférences de domaine permettant de cibler le centre d'intérêts de l'utilisateur,

. Historique de l'utilisateur : les interactions passées de l'utilisateur représentent une source pour prédire ses intentions et lui recommander des objets.

Les approches et techniques de la modélisation utilisateur peuvent être basées sur des modèles simples ou complexes dépendant de l'objectif final ou domaine d'application du système; un effort de standardisation pour la généralisation de tels systèmes afin de produire des *Shell* a toutefois été mené et semble donner une meilleure portée au devenir des méthodes de modélisation de l'utilisateur [KOB01].

5.2 Acquisition du profil

Un système de recherche d'information utilise une collection statique de document et des requêtes renouvelées que l'utilisateur soumet au système alors que dans le filtrage, le profil de l'utilisateur qui est une requête à long terme est relativement statique pendant que de nouveaux documents entrant dans le système (flot dynamique de documents).
Le profil utilisateur décrit ses préférences qui, comparées aux documents entrants permettent de sélectionner ceux susceptible de l'intéressé. Un cas typique de l'application d'une telle approche est utilisée dans la sélection de nouveaux articles parmi des milliers d'autres qui sont diffusés chaque jour [BAE99].
Le système de filtrage d'informations indique seulement à l'utilisateur les documents qui peuvent l'intéressé, la tâche consiste à déterminer lesquelles qui sont les plus pertinents est spécifiquement et strictement réservé à l'utilisateur, même si un classement des documents filtrés lui est donné.
Une autre variante de cette procédure est d'afficher le classement des documents filtrés pour qu'un utilisateur n'examine qu'un nombre très restreint de document s'il s'assure que les premiers classés ont des chances d'être les plus pertinents. Cependant même si le classement n'est pas présenté à l'utilisateur, le système de filtrage peut opérer un classement interne pour déterminer les documents les plus pertinents. N'importe quelle méthode peut être utilisée pour effectuer ce classement mais le modèle vectoriel est plus préféré de part sa simplicité.
Dans une tâche de filtrage, l'étape cruciale n'est pas le classement des documents par ordre de pertinence mais la construction d'un profil utilisateur qui reflète réellement ses préférences et intérêts.
Plusieurs approches pour la construction du profil utilisateur ont été proposées pour acquérir le profil de l'utilisateur et la qualité et la pertinence de cette acquisition, dépend étroitement des caractéristiques et fonctionnalités disponibles des systèmes et leurs capacités à extraire des informations qui décrivent son profil [BAE99].

5.3 Représentation du profil

Il n'y a pas de modèle spécifique dédié à la représentation du profil de l'utilisateur, on en cite principalement quatre types de représentation vectorielle, sémantique, connexionniste et multidimensionnelle.

• *Représentation vectorielle*

Ce type de représentation s'appuie généralement sur le modèle vectoriel [SAL71], le profil est représenté par un ou plusieurs vecteurs définis dans un espace de termes obtenus implicitement ou explicitement à partir de plusieurs sources d'information. Les coordonnées des vecteurs correspondent aux poids des termes dans le profil. L'utilisation de plusieurs vecteurs permet de prendre en compte la diversité des domaines d'intérêts ou évolution dans le temps.
Ce type de représentation offre l'avantage indéniable de la simplicité de mise en œuvre. Cependant les modèles proposés ne mettent pas en évidence ni la dimension liée au temps marquant l'évolution des profils, ni à l'organisation des informations pour hiérarchiser les centres d'intérêt.

- *Représentation sémantique*

La représentation sémantique met d'avantage en relief les relations de sens entre unités d'informations représentant le profil en apportant des solutions aux problèmes de dissémination et synonymie. La direction proposée dans ce contexte, est la construction hiérarchique de concepts plutôt qu'une liste de structures indépendantes [BOU05], à partir d'informations issues des fichiers *logs.* La hiérarchie peut rendre compte des niveaux de préférences de l'utilisateur et des associations latentes entre concepts et donc un raisonnement sémantique pour la dérivation du profil s'y prête aisément. Ce type de représentation utilise généralement des ontologies.

- *Représentation connexionniste*

C'est un type de représentation basé sur l'interconnexion de nœuds représentant les termes, préférences ou documents. Il offre le double avantage de la structuration et de la représentation associative permettant de considérer l'ensemble des aspects représentatifs du profil.

- *Représentation multidimensionnelle*

Différents travaux ont abordé cet aspect sans le couvrir dans son ensemble. Ainsi, les propositions de standards pour la sécurisation des profils ont défini des classes distinguant les *attributs démographiques* des utilisateurs (identité, données personnelles), *les attributs professionnels* (employeur, adresse, type) et *les attributs de comportement* (trace de navigation). [AMS99] propose cinq catégories d'informations pour les utilisateurs d'une bibliothèque digitale : *Données Personnelles* (identité), *Données Collectées* (contenu, structure et provenance des documents), *Données de Livraison* (moment et moyen de livraison), *Données de Comportement* (interactions de l'utilisateur avec le système), *Données de Sécurité* (conditions d'accès aux informations du profil). Ces tentatives de structurations sont louables mais insuffisantes pour couvrir le champ de la personnalisation. Par ailleurs, elles se contentent de catégoriser les informations de profil sans expliciter les corrélations qui existent entre elles. En effet, les attributs de comportement peuvent être corrélés aux attributs personnels ou professionnels. De même, les données de sécurité peuvent caractériser aussi bien les données personnelles que les données collectées.

Les données personnelles,
Le centre d'intérêt,
L'ontologie du domaine,
La qualité attendue des résultats délivrés,
La customisation
La sécurité et la confidentialité
Le retour de préférences (feedback)
Des informations diverses

Une autre dimensions est rajoutée pour les informations inclassables *DIVRS*, [BOK06].

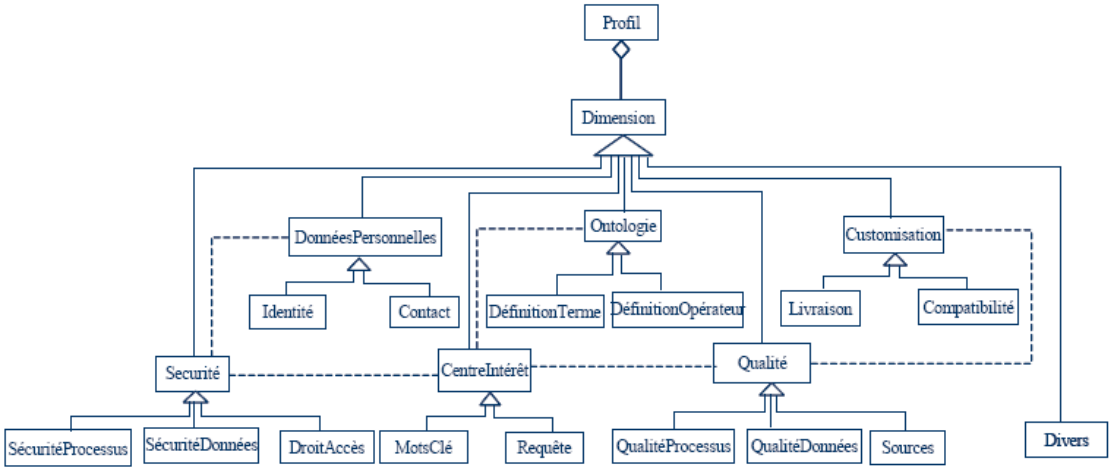


Fig I.5- Dimensions du modèle profil [BOU06]

- Optimisation du profil

Dans un système de filtrage les informations sont filtrées par rapport à un ensemble de mots clés qui composent le profil de l'utilisateur, pour cela plusieurs techniques sont déployées pour la construction automatique du profil de l'utilisateur, mais elles emploient généralement un grand nombre de mots clés pour le décrire. Dans une perspective d'optimisation en vue d'améliorer les performances des systèmes de filtrage, d'autres techniques préconisent la réduction de la dimension de ce profil par l'élimination d'un certain nombre de termes qu'il contient [BOU05].

Une des techniques est l'utilisation du modèle de réseau de neurones qui procède à l'analyse à travers un apprentissage automatique du profil utilisateur. Le réseau de neurone est entraîné pour identifier un ensemble optimal de mots clés pour classées les données et informations pertinentes pour l'utilisateur.

- Difficulté de modélisation

La description du profil utilisateur au départ est la tâche la plus ardue dans un système de filtrage d'information. A la base de ces informations le profil devra être modifié en conséquence.

Néanmoins deux risques s'apparentent au profil utilisateur, le premier est son opacité [PIN01], en effet au fur et à mesure de son évolution, on perd à quoi il correspond et quelle thématique il couvre. Le profil est continuellement alimenté et rectifié en fonction de son jugement de pertinence, les mots utilisé pour le décrire, sont privé de leur contexte et le risque est qu'une fois, accumulés, s'ils ne sont pas bien organisés peuvent devenir ininterprétables.

Le deuxième problème est la dégénérescence du profil [PIN01], si le système enrichit régulièrement le profil sans par ailleurs le rééquilibrer, ceci peut émousser la précision ou au contraire se polariser sur un aspect qui occulte les autres.

- Le profil et communauté

Un tel système de recommandation est fortement liée á la qualité de ces utilisateurs qui sont représentés par leurs profils et leurs contributions pour la formation des communautés (*regroupement des utilisateurs en fonction d'un critère*), les profils sont un facteur interactif, alors que les communautés sont considérées comme un facteur interne du système.

6. Évaluation de système de filtrage

Il existe plusieurs mesures d'évaluations pour un système de filtrage tels que la couverture, rappel, précision, le temps de réponse, effort fournie par l'utilisateur, présentation des résultats, etc. [NOU04].
La figure ci-dessous présente les partitions d'une collection pour un exemple de filtrage de documents.

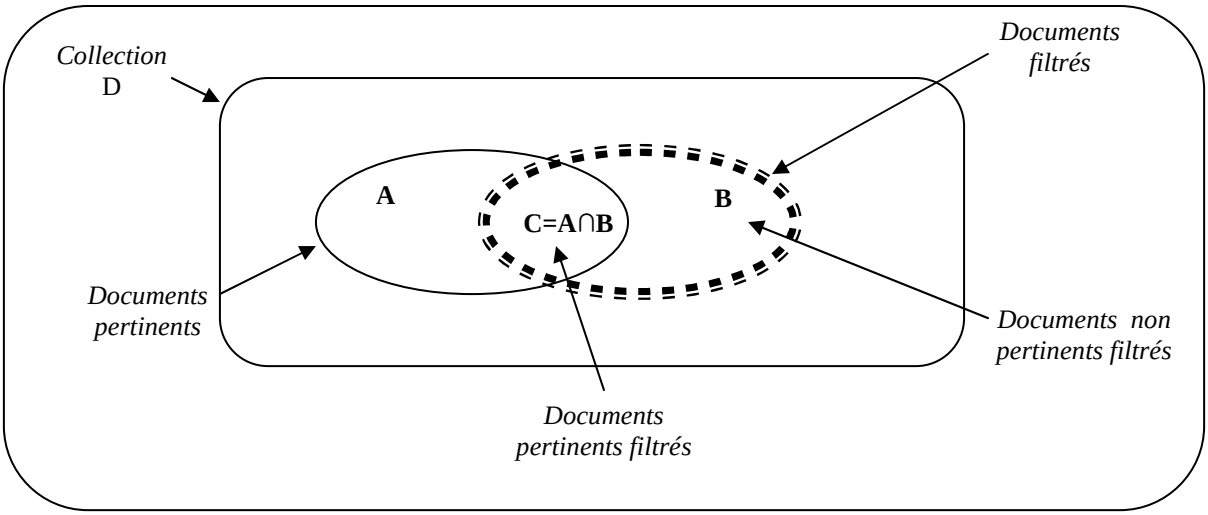


Fig I.6- Partition de la collection pour un SF

Précision : c'est la proportion des documents pertinents bien filtrés par le système par rapport au nombre total de documents filtrés par ce dernier.

$$\Pr écision = \frac{|C|}{|B|}$$

Rappel (recall) : c'est la proportion des documents pertinents bien filtrés par le système par rapport à ce qu'il aurait dû filtrer au meilleur des cas.

$$Rappel = \frac{|C|}{|A|}$$

Fallout : c'est la proportion des documents filtrés comme étant pertinent mais qui ne le sont pas, par rapport à l'ensemble des documents qui ne devaient pas être filtrés par le système.

$$Fallout = \frac{|\ B - C\ |}{|\ (B - c)\ | + |\ (D - A)\ |}$$

La précision et le rappel sont les fonctions à maximiser pour un système. Ils varient de manière inversement proportionnelle [PIL00].

7. Quelques systèmes de filtrage

GroupLens

C'est un système expérimental de l'université du Minnesota, les lecteurs sont appelés à noter les articles qu'ils lisent sur une échelle numérique de cinq niveaux, le principe de fonctionnement repose sur le calcul de corrélation entre les différents utilisateurs et identifie des groupes d'utilisateurs dont les intérêts sont semblables, et ensuite il emploie ces estimations pour prédire l'intérêt que porteront les lecteurs à chaque article [MIL97] , Avec GroupLens, les estimations sont réparties en plusieurs emplacements et son architecture est ouverte à la création de nouveaux clients de newsgroups et serveurs d'estimation qui emploieraient l'évaluation d'une manière différente [BER03].
GroupLens a aussi permis de démontrer que la consultation des estimations des autres utilisateurs ne constituait pas un risque de biais pour l'évaluation [PAL97].

En raison du grand nombre de différents documents, ce système dépend beaucoup du nombre de lecteurs et de leurs évaluations sur les mêmes documents [MAL95]. De plus, il souffre d'un problème de démarrage à froid [MIL97]. Beaucoup d'utilisateurs ont abandonné son utilisation ; ils avaient un grand nombre de documents à noter avant de commencer à recevoir des recommandations et donc à bénéficier du système (problème de motivation). En outre, les premiers utilisateurs ne recevaient pratiquement que des documents qu'ils avaient déjà lus et notés, en raison de la lenteur de l'apprentissage.

CiteSeer

Un système de recommandation de pages web qui utilise les bookmarks personnels et leur organisation en répertoires pour prédire et recommander des pages pertinentes [RUC97].
Le système apprend pour chaque page web quelles sont les différentes communautés ou groupes d'affinités qui s'y intéressent les classifier selon leurs intérêt contextuel et génère des recommandations. Les bookmarks offrent un mécanisme de collecte d'information sur les préférences directement géré par l'utilisateur. CiteSeer consulte les bookmarks de chaque utilisateur et mesure le degré de chevauchement (URL communs par exemple) de chaque répertoire avec les répertoires d'autres utilisateurs pour donner un poids additionnel aux URL.
Comme toute approche collaborative, la limite de CitSeer est l'incapacité de servir les premiers utilisateurs ou un utilisateur créant une nouvelle catégorie.

Referralweb

Referralweb [KAU87] se présente comme un système interactif pour la construction, visualisation et la recherche de réseaux sociaux sur le web. Une reconstruction manuelle de ces réseaux est certes possible mais risque d'être frustrante et coûteuse en termes de temps.
Les utilisateurs peuvent être appelés à introduire la liste de leurs collègues proches, ou encore, on peut analyser les entêtes des mails. Cette dernière solution n'est pas sans poser des problèmes de confidentialité et de sécurité évidents. Pour Referralweb, les données sont récupérées sur le web. Il utilise la co-occurrence de noms de personnes dans des fenêtres de proximité à partir des homes pages, des listes des coauteurs dans des publications et références à des papiers, les échanges d'enregistrements personnels dans les archives des newsgroups. Quand un utilisateur s'abonne pour la première fois à Referralweb, un moteur de recherche classique est utilisé pour retrouver les documents où une mention de son nom est faite.

Il est important de signaler que Referralweb ne remplace pas les moteurs de recherche génériques comme AltaVista ou google, mais sert à augmenter l'efficacité et la focalisation des sessions de recherche. Il permet aussi d'une appropriation des résultats par l'utilisateur, en ramenant des documents écrits par des personnes qui lui sont proches. D'un autre côté,

Referralweb cherche à découvrir des réseaux sociaux existants plutôt que d'offrir les outils pour créer de nouvelles communautés, et à la différence des autres systèmes de recommandation qui favorisent l'anonymat.
Referralweb est basé sur la connaissance des interlocuteurs et de la crédibilité qu'on leur porte, d'autre part ne demande pas à ses utilisateurs de saisir une liste de leurs collaborateurs mais se base sur des ressources disponibles au public sur le web[BER03].

SyCoFiD (outils CERIST)

Un prototype ou (Boite a outils), s'inscrit sous la catégorie du filtrage par contenu permet essentiellement de collecter et filtrer des doucement doté d'un interface utilisateur pour les profils, ainsi que l'échange de messages.

Fig I.7- l'interface SyCoFiD

CoCofil (*Community-Oriented Collaborative Filtering*)**:**
Plateforme ouverte d'essai, particulièrement orientée vers la communauté. En effet elle intègre des fonctionnalités destinées à mieux exploiter la notion de communauté d'utilisateurs. Elle a déjà servie de plateforme pour concevoir et valider des fonctionnalités qui par la suite ont été intégrées de façon sélective dans un portail communautaire destiné aux chercheurs.

Elle a été conçue d'une manière modulaire pour permettre l'ajout d'éventuelles fonctionnalités, en plus du stockage systématique d'un maximum d'informations même celles qui sont que rarement exploitées.

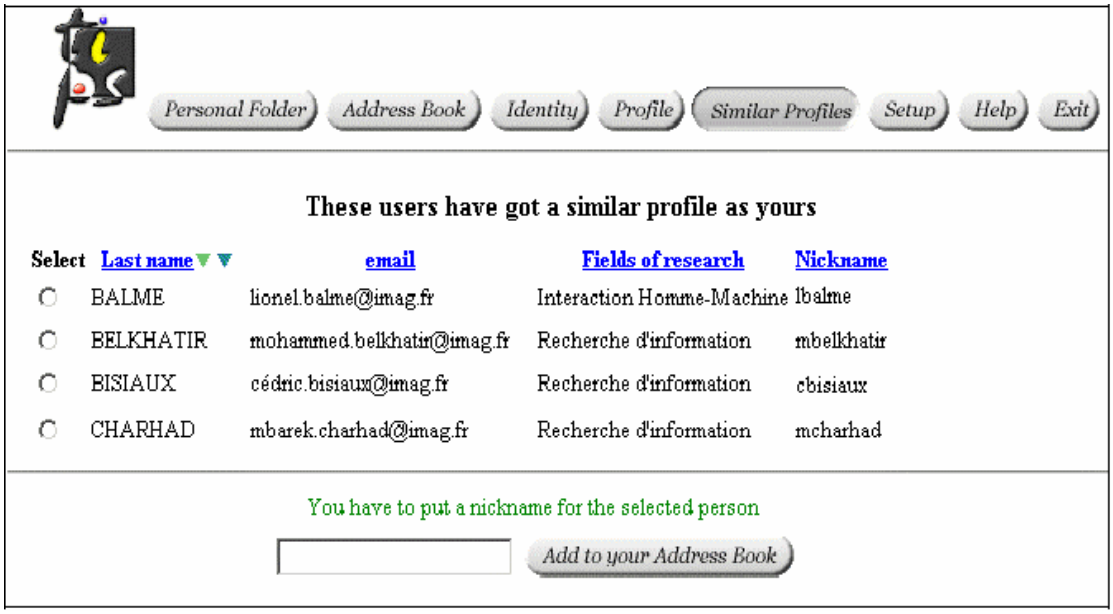


Fig I.8- Interface CoCofil2

A travers l'interface, l'utilisateur peut apercevoir les autres personnes du système, enregistre les adresses des utilisateurs intéressants, consulte le profil et visualise l'historique, en plus d'un Paramétrage personnaliser tels que le type d'affichage, la visualisation et l'option de l'anonymat, etc.

Ce système qui est basé sur la notion de multicritère dans la formation des communautés et l'intégration du filtrage actif pour échapper du démarrage à froid reste en phase de développement, les études techniques prouvent qu'il apporte une amélioration par rapport aux anciens systèmes.

8. Recherche et filtrage d'information

La recherche d'information RI et le filtrage de l'information FI sont les deux faces d'une même pièce.
La recherche inclut trois fonctionnalités principales qui sont l'organisation et l'indexation des documents qui sont des collections stable et la requête de l'utilisateur ainsi que la méthode de l'appariement entre les deux. Or, le filtrage d'information traite des documents entrants (flux entrant), et des profils utilisateurs (requête à long terme) ainsi la méthode employée dans l'appariement profils-documents. Le tableau ci-dessous éclaircit les points opposes de RI et FI.

Recherche de l'information	Filtrage de l'information
Concerne des usagers singuliers avec un objectif et requête.	Concerné par des usagers répétitifs par une ou plusieurs personnes avec des buts et des intérêts a long terme.
Travail sur l'adéquation requête (court terme) document	La requête est remplacée par le profil (long terme)
Collecte et organise les documents	La distribution des documents a des utilisateurs ou groupes.
Sélectionne des documents a partir d'une base classique de documents	Filtre ou élimine des documents à partir d'un flux dynamique de données.
L'interaction utilisateur–documents durant une session de recherche	Autorise le changement a long terme

Recherche dans les collections	Diffuse des informations

Tab I.2 - Comparaison entre recherche d'information et filtrage d'information.

9. Limites des Systèmes de filtrage Actuels

Comment décrire voire adapter au mieux les intérêts des utilisateurs (profils) et comment optimiser l'appariement entre ces profils et les informations disponibles, selon des méthodes efficaces sont les deux objectifs essentiels d'un système de filtrage.

Par ailleurs ces systèmes souffrent de quelques limitations en citant :

9.1 Difficulté de l'évaluation

Évaluer un système de recherche d'information pose problème dans la mesure où il est difficile d'y intégrer l'utilisateur, alors que c'est lui qui en dernier ressort, décide de la qualité du service rendu par le système. Ces difficultés se retrouvent bien évidemment avec les systèmes de filtrage collaboratif mais ils sont d'autant plus aigus que le service rendu par ce type de système qui doit s'évaluer au cours du temps tout au long de son exploitation.
En effet, pour l'utilisateur, le rapport entre le coût (son effort d'évaluation) et le bénéfice (les documents reçus automatiquement) varie au cours du temps, en particulier, au début de l'utilisation du système, ce rapport lui est souvent défavorable ce qui peut le décourager d'utiliser le système pour atteindre une phase plus favorable. La défection des utilisateurs pénalise alors l'ensemble des performances du système qui ne fonctionne bien qu'avec une participation active d'un nombre suffisant d'utilisateurs.

9.2 Démarrage à froid

A l'initialisation du processus de filtrage, on ne dispose d'aucune connaissance sur les documents à filtrer pour pouvoir construire une fonction de décision, ni pour bien pondérer les profils et les documents entrants. On distingue trois types de problèmes de démarrage à froid [BUR02].
– le démarrage à froid pour un nouveau système où les performances des systèmes sont très mauvaises en raison de l'absence d'informations sur lesquelles fonder le processus de filtrage.
– le démarrage à froid pour un nouveau document, c'est un problème spécifique à l'approche collaborative, pour laquelle les objets à recommander ne sont décrits que par les évaluations fournies par les utilisateurs.
– le démarrage à froid pour un nouvel utilisateur, le profil de l'utilisateur est inexistant et ses communautés sont encore inconnues, ce qui conduit à des recommandations de mauvaise qualité.

9.3 Adaptation du profil

Pour adapter les profils, l'approche la plus populaire est de les raffiner au fur et à mesure que les utilisateurs fournissent leurs évaluations [MLD03]. Néanmoins, un changement dans leur besoin d'information au niveau des centres d'intérêt n'est pas toujours bien pris en compte. C'est peut-être à cause de la lenteur du processus d'adaptation, il faut que l'utilisateur évalue parfois beaucoup de documents pour que se produise un changement significatif dans son profil.

9.4 Qualité de prédiction

Les documents et les profils sont représentés par des listes de mots pondérés. L'appariement document-profil consiste à mesurer une similarité. La décision quant à l'acceptation ou le rejet d'un document est assurée par une fonction de décision souvent de type seuil, la

détermination de ce seuil et les pondérations adéquates associées aux profils et aux documents sont les problèmes majeurs rencontrés ainsi, ces systèmes se basent sur les statistiques des termes et des documents dans les bases profils et les bases de référence pour estimer les valeurs de plusieurs paramètres de filtrage, or ces statistiques sont très variables au cours du filtrage et inconnues à l'initialisation du processus. L'insuffisance et l'imprécision de ces mesures conduisent à une exploitation des matrices creuses par le système, ce qui engendre une valeur de prédiction désagréable (qualité < 5 % pour Netflix et movielens).

9.5 Extension des domaines d'application

Les travaux expérimentaux actuels se trouvent notamment sur le plan économique tels que le *mailing list, Usnet News* et le filtrage des *e-mails*, etc. L'extension de ces systèmes vers d'autres domaines comme la documentation académique, la formation à distance (*e-learning*) permet d'inciter les développeur de trouver les outils robustes et les solutions adéquates pour faire face aux situations critiques.

9.6 Manque d'une couche sémantique

En tant qu'infrastructure, le web sémantique doit permettre d'utiliser des connaissances formalisées et des étiquettes en plus du contenu informel et multi-média actuel du web. L'objectif est de dépasser les modes classiques de recherche d'information, de pouvoir combiner des informations provenant de plusieurs sites ou sources pour répondre à des besoins élaborés (comme composer un séjour touristique en gérant à la fois le séjour, les transports, les activités et visites, etc.), de s'adapter aux préférences des utilisateurs, de leur proposer des informations pertinentes plutôt que des les aider à les chercher, etc. Cette infrastructure doit permettre d'abord de localiser, d'identifier et de transformer des ressources du Web de manière robuste et valide, tout en étant accessible à une plus grande diversité d'utilisateurs [BOU07].

Cette infrastructure permet d'ajouter un raisonnement dans la recherche en intégrant des métas donnés (des annotations) dans les profils utilisateur et dans la base des documents, ces annotations permettant de cibler l'information recherchée et augmenter le facteur de pertinence.

Le problème de l'hétérogénéité sémantique des données rend le processus de recherche en particulier le filtrage complexe ce qui implique la diminution des recommandation et la pertinences des documents.

Conclusion

Dans ce chapitre on a fait un tour d'horizon sur les systèmes de recherche en particulier les systèmes de filtrage et leurs formes.

En effet, le filtrage collaboratif peut s'avérer utile pour conseiller les utilisateurs sur certains documents multi formats qu'ils n'ont pas lus, et dont on sait qu'ils intéressent son groupe.

Il est possible de regrouper les connaissances prises en compte dans un système de filtrage les connaissances sur les documents (index), les connaissances sur les utilisateurs (profils ou modèles utilisateurs) et enfin, les connaissances sur les concepts du domaine de l'application. Ces dernières peuvent servir de référence pour reformuler les recommandations pour utilisateurs.

Les systèmes de filtrages actuels souffrent de manque de précision dû a la non prise en compte du niveau sémantique, dans ce travail nous souhaitons améliorer la génération de la prédiction en introduisant l'aspect sémantique, l'objectif est d'apporter une amélioration à l'efficacité du filtrage en mécanisme de source d'information et en tirant profit de l'infrastructure web sémantique.

CHAPITRE II
WEB SÉMANTIQUE ET FILTRAGE D'INFORMATION

Dans ce chapitre, nous introduisons les différentes notions liées au web sémantique et à la représentation de la similarité entre ressources selon des normes définies par W3C, La représentation de l'information sémantique peut cependant être faite de plusieurs manières alternatives, dans des bases de données relationnelles ou autres, nous nous focalisons sur l'effet de ces données sémantiques pour booster les SFC qui marquent des limitations critiques telles que le manque de données d'évaluation et le démarrage à froid.

1. Définition et propriétés

La notion de web sémantique prend de plus en plus d'importance et devient l'une des voies d'évolution les plus prometteuses du Web, tel qu'on le connaît aujourd'hui. La définition la plus connue est celle donnée par Tim Berners-Lee[1].

« Le Web sémantique est une extension du Web actuel, dans laquelle l'information reçoit une signification bien définie, améliorant les possibilités de travail collaboratif entre les ordinateurs et les personnes ».

Cela semble donc signifier, dans l'esprit des auteurs, que l'information n'a pas de signification bien définie dans le Web actuel et la majeure partie de l'information est sous forme textuelle, très peu structurée et donc inutilisable pour faire des traitements de calcul ou d'inférences. Il est pourtant bien évident que l'information disponible sur le Web actuel a une signification, mais qu'elle n'est accessible aujourd'hui qu'à des lecteurs humains.

En première approximation, le but du web sémantique est de développer un web dont le contenu s'adresse, au moins pour partie, aux machines afin qu'elles puissent aider les utilisateurs humains [CHA03]. Si l'on cherche à préciser, un tel web doit doter ses ressources (documents, service…) d'annotations dont le but n'est pas d'assurer l'affichage des documents mais l'appréhension de son contenu par divers outils logiciels. Le web sémantique doit donc être une infrastructure juxtaposant au web actuel. Des documents structurés par des langages pour exprimer la connaissance, pour décrire les relations entre les connaissances, pour décrire les conditions d'utilisation et pour décrire les garanties et les modes de payement, et des dispositifs permettant de trouver les ressources.

Les recherches actuellement réalisées s'appuient sur un existant riche venant, par exemple, des recherches en représentation ou en ingénierie des connaissances. Mais leur utilisation et leur acceptation à l'échelle du (ou d'une partie du) Web posent de nouveaux problèmes et défis : changement d'échelle dû au contexte de déploiement, le Web et ses dérivés (intranet, extranet), nécessité d'un niveau élevé d'interopérabilité, ouverture, standardisation, diversités des usages, distribution bien sûr et aussi impossibilité d'assurer une cohérence globale. Comme l'écrit, en substance, Tim Berners-Lee, *"le Web sémantique est ce que nous obtiendrons si nous réalisons le même processus de globalisation sur la représentation des connaissances que celui que le Web fit initialement sur l'hypertexte".*

Nous montrerons également le rôle que peuvent jouer cette nouvelle l'infrastructure dans le processus de filtrage d'information et voir l'impacte majeur pour la pertinence d'information et la qualité de ce type de systèmes.

D'après W3C[2] :Le Web Sémantique est une vision, l'idée que les données sur le Web soient définies et liées de manière à pouvoir être utilisées par des machines non seulement pour des fins d'affichage, mais pour l'automatisation, l'intégration et la réutilisation sur des plates-formes variées.

---

[1] http://www.w3.org/2001/sw/

[2] World Wide Web Consortium

La finalité du Web sémantique sera donc de hiérarchiser toutes les informations présentes sur la toile dans le but de pouvoir effectuer des recherches précises complexes et rapides.

Les recherches actuelles sur le WS proposent de s'appuyer sur des techniques de représentation de connaissances (formalisme et raisonnement) pour munir l'information contenue dans les ressources Web d'une sémantique.

2. Fondations du WS

Le WS n'est pas construit à partir de zéro, le Web actuel est déjà une base immense des connaissances et des informations où les informations sont sauvegardées, organisées et représentées de façon passive, non structurée, arbitraire et ad hoc. Le Web sémantique hérite du web actuel des connaissances mais il permet aussi à des machines d'accéder à cette base et de l'exploiter.
Le Web sémantique est donc une extension du Web actuel, la figure montre en image les extensions du Web sémantique par rapport au Web actuel [BAC06].

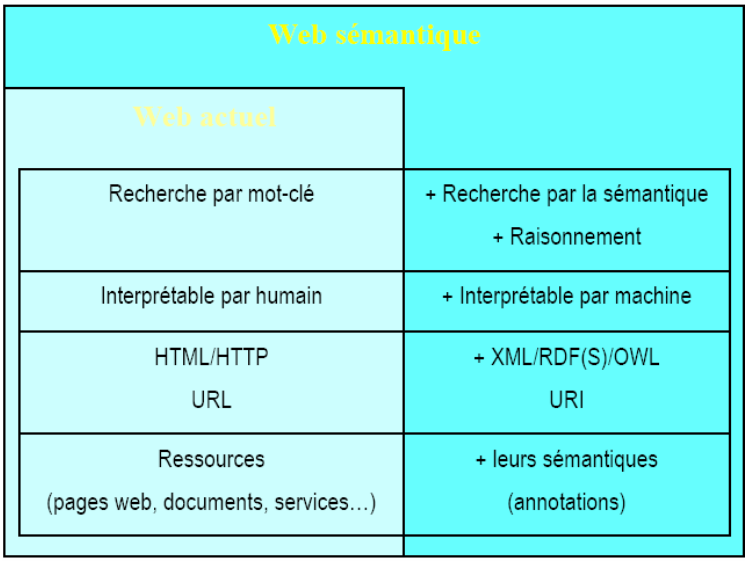


Fig II.1 -Le Web actuel et son extension, le Web sémantique

La réalisation de cette nouvelle infrastructure [BOU07] du Web s'appuie sur l'utilisation de :
. **Méta-données** : par définition des données sur des données elles complètent donc l'information sur les données à un niveau d'abstraction supérieure, elles peuvent être structurées afin de décrire une ressource quelconque et rajoutent un sens aux contenus afin de favoriser leur exploitation par des agents logiciels.

. **Des ontologies :** permettant de la spécification des concepts voire aussi des axiomes liés à un certain domaine, modélisent les connaissances nécessaires à la description et au traitement d'un ensemble de ressources.

. **Des langages** : différents langages pour décrire, exploiter et raisonner sur les contenus des ressources ainsi qu'une représentation de connaissances afin d'exprimer les ontologies et décrire les annotations.

. **Des moteurs de raisonnement :** encapsulés dans des systèmes de requêtes et permettant d'inférer sur les annotations d'après les axiomes déclarés dans les ontologies, afin d'interroger le Web et agir sur les réponses obtenues.
Le Web informel est déjà disponible, c'est la formalisation qui fait le WS.

3. Architecture du WS

Dans cette section, nous présentons les composants de base sur lesquels on construit le Web de demain, (Figure II.2) [PAT06].

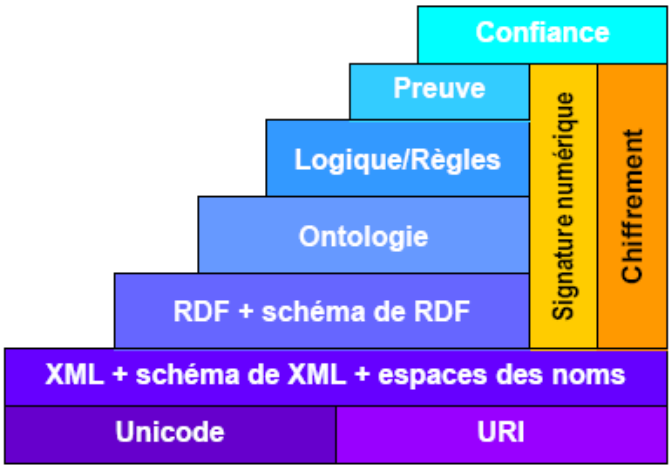


Fig II.2 - Architecture du web sémantique.

3.1 URI

Le premier facilitateur pour les technologies du Web sémantique est l'URI[3] (Uniform Resource Identifier - identifiant uniforme de ressource). Une URI est une chaîne qui identifie une ressource ou un concept et employée pour désigner le nom ou l'adresse (ou les deux) d'une ressource dans le Web sémantique sans distinction d'une ressource physique (donc sa représentation est récupérable via l'internet telle qu'une page web, un service localisé sur un serveur…) ou d'une ressource abstraite (un livre particulier, une idée…).

Remarque **:** un URL (Uniform Resource Locator) sont des URI qui permettent de localiser (d'accéder) à la ressource en utilisant ftp, mailto, gopher, etc. Ces ressources sont toujours disponibles sur le web ce qui ne pas toujours le cas pour les URIs.

Quelques exemples d'URI :

- http://www.inria.fr/acacia/index.htm (pour une page web)
- http://www.inria.fr/acacia/OntologyMatching.pdf (pour un article)
- rstp://www.video.com/france.rm (pour une vidéo)
- urn:issn:2242-4157 (pour un livre)
- tel:+33-497-15-53-17 (pour un numéro de téléphone)

[3] http://www.gbiv.com/protocols/uri/rfc/rfc3986.html

3.2 XML

XML[4] (e**X**tensible **M**arkup **L**anguage), fournit une syntaxe pour les documents Structurés dans une structure hiérarchique mais n'impose aucune contrainte sémantique à la signification de ces documents. Un langage à balises combine le texte et les informations supplémentaires (méta-données) sur le texte Développé par le W3C et standardisé en 1998. Les informations supplémentaires telles que la structure, la police, la couleur…, sont exprimées en utilisant des balises, qui sont mélangées avec le texte à présenter. En plus, le langage XML permet aux utilisateurs de définir eux-mêmes leurs balises.

XML, langage simple pour la création de documents auto descriptifs. Ces caractéristiques résident dans son accessibilité, internationalisation, indépendance par rapport au mode d'accès et évolution très rapide.

- Syntaxe XML

. Eléments

L'élément est le concept principal d'un document XML.
C'est un triplet <balise d'ouverture, contenu, balise de fermeture>.

<enseignant> Patrice Buche </enseignant>

Un élément peut être vide :

<enseignant> </enseignant> .

. Attributs

Un attribut permet d'associer des propriétés à un élément.

<enseignant name="Patrice Buche" tel="+33 1 44 08 16 75"/>

<commande noCde="123" client="Dupont">

<item itemNo="a100" quantité="1"/>

<item itemNo="a102" quantité="3"/>

</commande>

. Documents XML bien formés

Un document XML est bien formé s'il est syntaxiquement correct. Les règles principales sont :
· Il a un seul élément racine.
· Chaque élément a une balise ouvrante et fermante.
· Les attributs d'un élément doivent avoir un nom unique.

. Structure d'arbre

Un document XML bien formé a une structure d'arbre étiqueté ordonné :
· Il n'y a qu'un élément racine.
· Il n'y a pas de cycle.
· Chaque noeud a un seul parent (sauf le noeud racine).
· Chaque noeud a une étiquette.
· L'ordre des éléments est important (par contre, l'ordre des attributs n'a pas d'importance).

Soit l'information: BENAMER Mohamed, 12, Place des Martyrs, Alger. 021 15 37 59.
Cette information est structurée en XML comme suit:

<?xml version="1.0" encoding="utf-8"?>

---

[4] http://xmlfr.org.

<personne>

<prénom>Mohamed</prénom>

<nom>BENAMER</nom>

<adresse>12, Place des Martyrs, Alger </adresse>

<tel> 021 15 37 59</tel>

</personne>

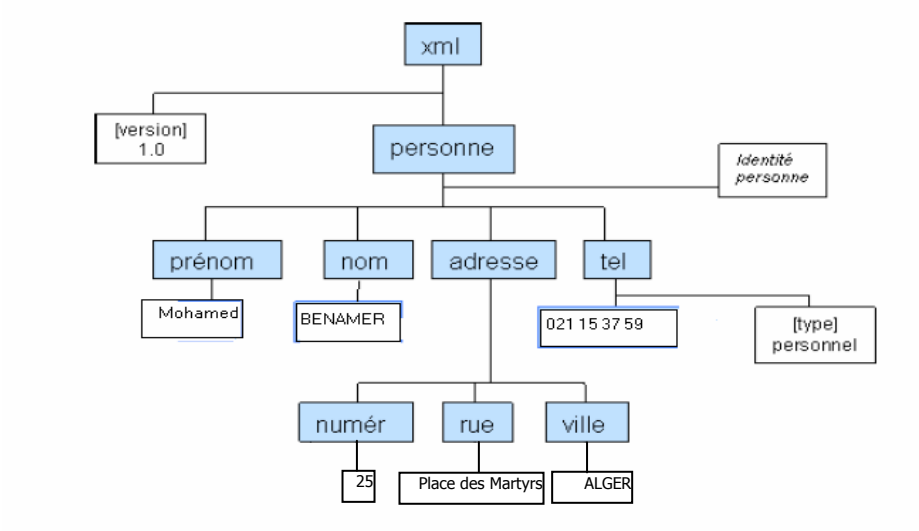


FIG II.3 – Information structurée

. Structuration des données

Un document XML bien formé est valide s'il respecte la structure définie par une DTD ou un schéma XML.

. DTD

L'élément

<enseignant>

<name>Patrice Buche</name>

<tel>+33 1 44 08 16 75</tel>

</enseignant>

a pour DTD

<!ELEMENT enseignant (name, tel)>

<!ELEMENT name (#PCDATA)>

<!ELEMENT tel (#PCDATA)>

. Schéma XML

Langage plus riche que les DTD pour décrire la structure d'un document XML.
XML schéma respecte une syntaxe XML : réutilisation de la technologie XML (Analyseurs, éditeurs, ...), possibilité de réutiliser, d'étendre ou de restreindre un schéma existant.

Déclaration du schéma EnseignantType (fichier enseignant.xsd)

<?xml version="1.0"?>

<xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema" version="1.0">

```
<xsd:element name="Enseignant" type="EnseignantType"/>
<xsd:complexType name="EnseignantType">
<xsd:sequence>
<xsd:element name="Prenom" type="xsd:string" minOccurs="0"
maxOccurs="unbounded"/>
<xsd:element name="Nom" type="xsd:string"/>
</xsd:sequence>
<xsd:attribute name="title" type="xsd:string" use="optional"/>
</xsd:complexType>
</xsd:schema>
```

. Espace de noms

Un document XML peut utiliser plusieurs DTD ou schémas, des conflits sur les noms utilisés peuvent apparaître. Si un même élément est défini de manière différente dans deux schémas, l'analyseur doit savoir quel schéma appliquer. Pour cela, les noms sont préfixés par un nom de schéma.

Un espace de noms est une collection de noms (d'éléments et de propriétés) identifiée par une URI.

Les espaces de noms sont déclarés dans un élément et peuvent être utilisés dans cet élément ou dans ses fils (éléments ou attributs) [PAT06].

Un schéma décrivant les enseignants de l'INA P-G (xmlschema2.xsd)

```
<?xml version="1.0"?>
<xsd:schema xmlns:ens=http://www.inapg.fr/Enseignant
xmlns:xsd=http://www.w3.org/2001/XMLSchema
targetNamespace="http://www.inapg.fr/Enseignant" version="1.0">
<xsd:element name="Enseignant" type="ens:EnseignantType"/>
<xsd:complexType name="EnseignantType">
<xsd:sequence>
<xsd:element name="Prenom" type="xsd:string" minOccurs="0"
maxOccurs="unbounded"/>
<xsd:element name="Nom" type="xsd:string"/>
</xsd:sequence>
<xsd:attribute name="title" type="xsd:string" use="required"/>
</xsd:complexType>
</xsd:schema>
```

Interrogation de données XML

Il existe plusieurs propositions de langages d'interrogation de documents XML (XQL, XML-QL, XQUERY, XPATH). Le concept central de ces langages est l'expression de chemin qui indique comment on peut atteindre un noeud dans la représentation arborescente d'un document XML.

XPATH est un langage qui permet d'adresser une partie de document XML non seulement avec l'objectif d'interroger, mais aussi pour pouvoir transformer un document XML (avec XSL).

3.3 RDF

RDF (Resource Description Framework) est un modèle de méta-données pour référencer des objets (ressources) et comment ils sont reliés l'un à l'autre. Dans ce modèle, les ressources sont identifiées (référencées) par les URIs et nous pouvons faire des déclarations à propos de ces ressources en employant des expressions sous forme « sujet – prédicat – objet » appelées des triplets. Le sujet est la ressource, la « chose » à décrire, le prédicat est un attribut ou un aspect de cette ressource qui exprime un lien ou une relation entre le sujet et l'objet, l'objet est l'objet de la relation ou de la valeur de ce prédicat.
Les triplets RDF peuvent se décrire en XML.

Par exemple : BENAMER Mohamed est le propriétaire de la page Web *http://www.ustomb.dz/infop/Mohamed.htm*

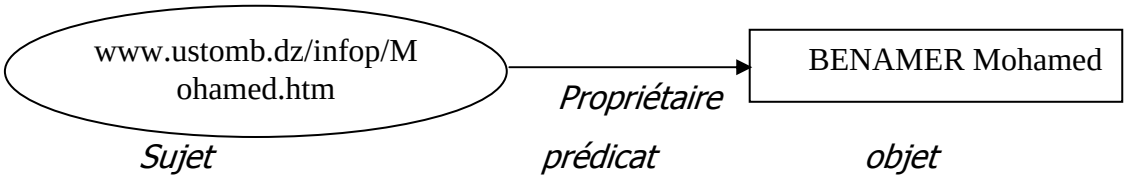


FIG II.4 - Un exemple d'un modèle RDF

· La représentation choisie pour les traitements automatiques est exprimée en syntaxe XML.
· Un document RDF est un arbre XML dont l'élément racine a pour nom rdf:RDF.

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:monDom="http://www.usto.dz/mon-rdf-ns#">
<rdf:Description rdf:about="http://www.ustomb.dz/infop/Mohamed.htm">
<monDom:proprietaire>
BENAMER Mohamed
</monDom:proprietaire>
</rdf:Description>
</rdf:RDF>
```

3.4 RDF Schéma

RDF Schema[5] est un langage pour décrire des vocabulaires, des propriétés et des classes de ressources dans le modèle RDF. RDF Schema est une extension sémantique de RDF. Il fournit

[5] http://www.w3.org/TR/rdf-schema/

des mécanismes pour décrire des groupes de ressources similaires (classes) et des relations entre ces ressources (propriétés). Les descriptions de vocabulaire de RDF Schema sont écrites en RDF en utilisant les termes (primitives) décrient dans la spécification du schéma RDF. La combinaison de RDF Schema et RDF est souvent référencée par RDF(S). Autrement dit RDF(S) fournit un moyen pour décrire les types des ressources et leurs caractéristiques.

Dans l'exemple suivant, nous employons des primitives du schéma RDF (tels que rdfs:Class, rdfs:label…) pour définir la classe « Article » (un groupe d'articles scientifiques) et décrire des caractéristiques de cette classe, donc des caractéristiques des instances (ressources) appartenant à cette classe, telles que leur titre, leur nombre de pages. Toutes les descriptions sont écrites en RDF.

```
<rdf:RDF
xmlns:rdf ="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:xsd ="http://www.w3.org/2001/XMLSchema#"
xml:base ="http://www.inria.fr/acacia/exemple#">
<rdfs:Class rdf:about="Article">
<rdfs:label xml:lang="en-US">l'article scientifique</rdfs:label>
</rdf:Class>
<rdfs:Property rdf:about="title">
<rdfs:label xml:lang="en-US">titre de l'article scientifique</rdfs:label>
<rdfs:domain rdf:resource="Article"/>
<rdfs:range rdf:resource="&xsd;string"/>
</rdf:Class>
<rdfs:Property rdf:about="pageNum">
<rdfs:label xml:lang="en-US">nombre de pages</rdfs:label>
<rdfs:domain rdf:resource="Article"/>
<rdfs:range rdf:resource="&xsd;integer"/>
</rdf:Class>
</rdf:RDF>
```

Une des caractéristiques les plus importantes RDF Schema est que nous pourrons définir les liens de « subsomption » entre des classes et des relations en employant les primitives rdfs:subClassOf et rdfs:subPropertyOf. Ce sont les liens de «spécialisation » ou « is_a » qui permettent aux classes et aux relations d'hériter des caractéristiques définies dans des classes (ou des relations) parentes. Cela permet des raisonnements dans RDF(S).

3.5 OWL

OWL[6] (Web Ontology Language) est un langage pour représenter des ontologies dans le Web sémantique. C'est une extension du vocabulaire de RDF(S). OWL est dérivé du langage

[6] http://www.w3.org/TR/owl-ref/

d'ontologie DAML+OIL[7]. En comparaison avec RDF(S) pour décrire des propriétés et des classes, OWL permet en plus d'exprimer, entre d'autres, des relations entre des classes (telles que la disjonction), la cardinalité (par exemple "exactement un"), l'égalité, plus des types des propriétés (propriétés d'objet ou d'annotation…), des caractéristiques des propriétés (par exemple la symétrie, la transitivité), et des classes énumérées. Une ontologie est capable de décrire des rapports entre des types de choses mais ne contient pas n'importe quelles informations indiquant comment employer ces rapports dans les calculs [BAC06].

Voici un exemple d'une ontologie en OWL :

```
<rdf:RDF
xmlns:owl ="http://www.w3.org/2002/07/owl#"
xmlns:rdf ="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:xsd ="http://www.w3.org/2001/XMLSchema#"
xmlns ="http://www.inria.fr/acacia/exemple/animals.owl#"
xml:base ="http://www.inria.fr/acacia/exemple/animals.owl#" >
<owl:Class rdf:ID="Animal">
<rdfs:label>Animal</rdfs:label>
<rdfs:comment>
This class of animals is illustrative of a number of ontological idioms.
</rdfs:comment>
</owl:Class>
<owl:Class rdf:ID="Person">
<rdfs:subClassOf rdf:resource="#Animal"/>
<rdfs:subClassOf>
<owl:Restriction>
<owl:onProperty rdf:resource="#hasParent"/>
<owl:allValuesFrom rdf:resource="#Person"/>
</owl:Restriction>
</rdfs:subClassOf>
</owl:Class>
<owl:ObjectProperty rdf:ID="hasAncestor">
<rdf:type rdf:resource="&owl;TransitiveProperty"/>
<rdfs:domain rdf:resource="#Animal"/>
<rdfs:range rdf:resource="#Animal"/>
</owl:ObjectProperty>
<owl:ObjectProperty rdf:ID="hasParent">
<rdfs:subPropertyOf rdf:resource="#hasAncestor"/>
```

[7] http://www.daml.org/2001/03/daml+oil-index

</owl:ObjectProperty>

</rdf:RDF>

Trois niveaux de langage OWL d'expressivité décroissante ont été définis :

· **OWL Full :** inclut toutes les primitives OWL qui peuvent être combinées avec toutes les primitives RDF et RDFS. Il a l'avantage de la compatibilité complète avec RDF/RDFS mais l'inconvénient d'avoir un niveau d'expressivité qui le rend indécidable.

· **OWL DL (Description Logic) :** définit des restrictions sur la manière dont les primitives OWL et RDF/RDFS peuvent être combinées. Ce niveau de langage est décidable, il a l'inconvénient de ne pas être complètement compatible avec RDF. Un document RDF devra donc parfois être étendu sur certains aspects ou restreint sur d'autres pour être un document OWL-DL légal.

· **OWL Lite :** est une version limitée d'OWL-DL qui en facilite son utilisation et son implémentation. Par exemple OWL-Lite n'inclut pas la disjonction entre classes et les cardinalités arbitraires.

Dans l'architecture du Web sémantique (FigureII.2, inspirée de l'architecture proposée par Berners-Lee, nous voyons les liens entre des éléments du Web sémantique présentés ci-dessus. L'URI, l'Unicode, XML sont considérés comme des briques de base, sur lesquelles reposent des langages tels que RDF(S) permettant de décrire des « choses » (objets, concepts…) et leurs types ; puis des langages tels que OWL pour exprimer des relations entre des types des choses mais sans spécifier comment exploiter ces informations dans les calculs. En combinant avec la logique, les règles, des assertions autour du Web (stockées par des langages des couches inférieures tels que RDF(S), OWL) peuvent être employées pour déduire de nouvelles connaissances.
Au-dessus de la couche de logique, c'est la couche de preuve, ici, nous avons un langage universel pour représenter des preuves et qui permet à différents systèmes de partager leurs connaissances « originales » ou déduites. La provenance (l'origine) des connaissances, des données, des ontologies ou des déductions est authentifiée et assurée par des signatures numériques, dans le cas où la sécurité est importante ou le secret nécessaire, le chiffrement est employé. Enfin, le but final vers lequel nous nous orientons est la confiance, un Web sémantique et fiable où nous pouvons effectuer plusieurs tâches complexes en sûreté [BAC06].

4. Propriétés du WS

On peut retenir trois propriétés importantes du Web Sémantique :

Formalisé : le Web informel est déjà disponible, c'est la formalisation qui fait le Web Sémantique. En effet, le Web actuel est composé principalement de pages HTML écrites à la main ou générées automatiquement pour un traitement humain.

Ouvert : le Web ne peut pas être sémantique dans une organisation et syntaxique à l'extérieur. Ce qui est important est que ce qui peut circuler soit formalisé. C'est à cette condition qu'il sera possible aux machines d'exploiter le Web Sémantique et, en particulier, de mettre en relation les ressources.

Interopérable : car il ne s'agit pas de communiquer dans un langage formalisé mais de savoir manipuler correctement cette connaissance formalisée.

5. Ontologies

5.1 Présentation

Gruber [GRU93] a définit l'ontologie comme la « spécification formelle d'une conceptualisation ». L'ontologie fournit la « compréhension commune des définitions de termes ou du vocabulaire dans un domaine défini ».

Aujourd'hui, les travaux sur les ontologies constituent un aspect de recherche important en informatique (Gómez-Pérez et al. 2004; Bouchard & Obaid, 2005). Ces travaux portent sur différents domaines incluant l'ingénierie des connaissances et l'intelligence artificielle. Les applications informatiques liées aux ontologies sont multiples : gestion des connaissances, traitement du langage naturel, commerce électronique, etc.
L'ingénierie ontologique est une discipline de l'informatique référant aux activités reliées au processus de développement des ontologies ainsi qu'aux méthodes, outils et langages pour développer ces ontologies. Plusieurs ontologies ont déjà été développées avec différentes méthodes.
En effet, une ontologie n'est pas opérationnelle, au sens où elle n'inclut pas de mécanismes de raisonnement, puisqu'elle doit justement être indépendante de tout objectif opérationnel. Le langage cible doit donc permettre de représenter les différents types de connaissances (connaissances terminologiques, faits, règles et contraintes) et de manipuler ces connaissances à travers des mécanismes adaptés à l'objectif opérationnel du système conçu. Ce processus de traduction est appelé *opérationnalisation*.

5.2 Construction d'une ontologie

Le processus général de construction d'ontologies peut donc être découpé en trois phases (figure II.5) [FRE04] :

. **La conceptualisation** : identification des connaissances contenues dans un corpus représentatif du domaine considéré. Ce travail doit être mené par un expert du domaine (ou un groupe d'experts), assisté par un ingénieur de la connaissance, apportant son expertise des paradigmes de représentation des connaissances en machine pour aider à la structuration des connaissances
. **L'ontologisation** : formalisation, autant que possible, du modèle conceptuel obtenu à l'étape précédente. Une part des connaissances du domaine peut, à ce niveau, être abandonnée, du fait de l'impossibilité de lever certaines ambiguïtés ou du fait des limitations de l'expressivité du langage de représentation d'ontologie utilisé. Ce travail doit être mené par l'ingénieur de la connaissance expert du modèle formel de représentation de l'ontologie, assisté de l'expert du domaine.

. **L'opérationnalisation** : transcription de l'ontologie dans un langage formel et opérationnel de représentation de connaissances.

Il est à noter que ce processus n'est pas linéaire et que de nombreux aller-retour sont a priori nécessaires pour bâtir une ontologie adaptée aux besoins. Remarquons pour finir que ce modèle de construction d'ontologie est ascendant, c'est-à-dire que l'on part des connaissances à représenter, pour aboutir à une représentation formelle. Mais une construction descendante est possible, qui consiste à choisir un modèle opérationnel de représentation, en fonction de l'objectif d'utilisation de l'ontologie, puis à instancier ce modèle avec les connaissances du domaine.

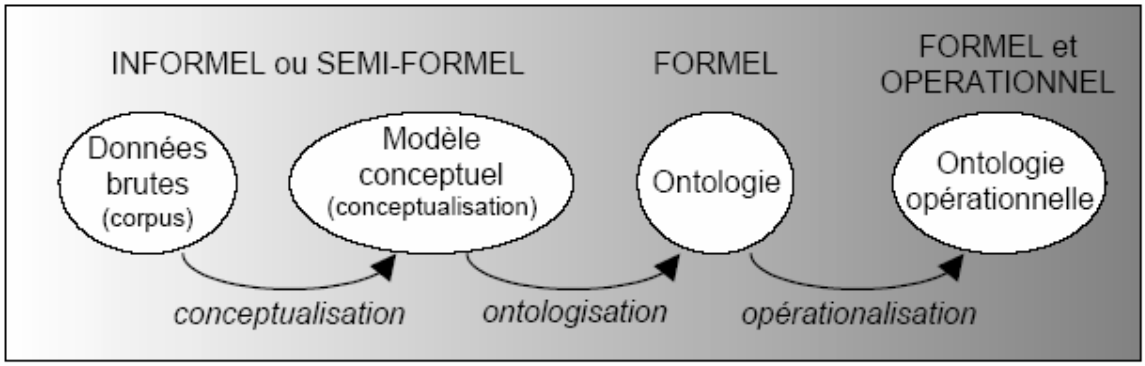


Fig II.5 - Construction d'une ontologie

5.3 Outils de développement

Les outils de développement d'ontologies qui existent sur le marché aujourd'hui sont divers et variés.
Dans cette section nous passons en revue quelques principaux outils disponibles.

. OILED

OILED est un éditeur graphique d'ontologies développé en java, qui utilise le formalisme DAML+OIL [BEC01]. Les cadres sont également exploités pour la modélisation. Le RDF Schema de DAML+OIL est utilisé pour le chargement et le stockage des ontologies. L'outil dispose de mécanismes pour la classification et le contrôle de la cohérence des ontologies. La version 3.4 de OILEd est gratuite et disponible sur le site http://oiled.man.ac.uk/.

. ONTOEDIT

OntoEdit [SUR02] est un environnement graphique d'ingénierie des ontologies (développement et maintenance), disposant d'un modèle ontologique interne et intègre, dans sa version commerciale, un serveur destiné à l'édition d'une ontologie par plusieurs utilisateurs ainsi qu'un plug-in permettant le test de la cohérence d'une ontologie.
Il est disponible en versions gratuite et professionnelle.

. ONTOLINGUA

Le serveur Ontolingua[8] [FAR96] rassemble un ensemble d'outils et services pour supporter la construction d'ontologies partagées entre différents groupes. Il a été développé par le Knowledge Systems Laboratory (KSL) de l'université de Standford. Il supporte plusieurs langages et dispose de traducteurs permettant de passer de l'un à l'autre. Il est aussi doté d'une bibliothèque d'ontologies accessible à distance ou localement via des éditeurs d'ontologies ou des applications.

. PROTÉGÉ-2000

Protégé-2000 [NOY01] est un outil d'acquisition et gestion de connaissances développé à l'Université de Stanford. Il est gratuit et disponible à l'adresse http://protege.stanford.edu. Il est doté d'un environnement graphique et interactif pour la conception d'ontologies et le développement de bases de connaissances. C'est un système ouvert et modulaire.

. WEBONTO

[8] http://www.ksl.stanford.edu/software/ontolingua/

WebOnto32 [DOM98] est un outil de création, visualisation et édition coopérative d'ontologies, développé par le KMI (Knowledge Media Institute - Open University,England). Il est doté d'une interface graphique et permet la modélisation de tâches. Le langage de modélisation des connaissances OCML (Operational Conceptual Modelling Language) est utilisé pour la spécification des ontologies. Il dispose d'un module de contrôle et de cohérence. Il est gratuit et disponible à l'adresse http://webonto.open.ac.uk.

5.4 Applications de l'ontologie

L'ontologie constitue un modèle d'un domaine et fournit un vocabulaire pour spécifier les besoins des applications et guide le développement de systèmes opérationnels. L'idée de base est d'utiliser l'ontologie pour permettre à des humains ou à des applications différentes d'accéder, par le partage ou l'échange, à des sources hétérogènes. Dans ce cadre l'ontologie facilite l'interopérabilité et la réutilisation de connaissances [MAR05].
L'utilisation de l'ontologie permettre à des humains ou à des applications différentes d'accéder par le partage ou l'échange à des sources hétérogènes. Dans ce cadre l'ontologie facilite l'interopérabilité et la réutilisation de connaissances. La recherche d'information est fondée sur des concepts afin de retrouver des documents pertinents. L'ontologie est utilisée par un moteur de recherche pour accéder à des ressources dans un répertoire.
Les connaissances ontologiques permettent de représenter le sens de la requête et d'effectuer des inférences sur les informations en décrivant le contenu de ressources et ainsi permettant l'amélioration de la qualité de la recherche.

Dans le cadre de filtrage d'information, il existe des ontologies pour la description des utilisateurs et de communautés comme : vCard (VisitCard), FOAF (Friend of a Friend), SIOC (Semantically Interlinked Online Communities), etc.

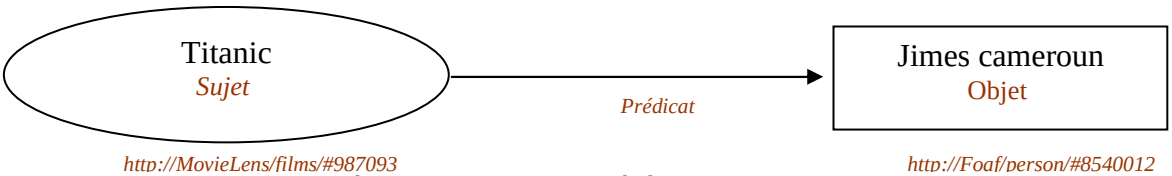
6. La description sémantique du profil

6.1 Les méta-données

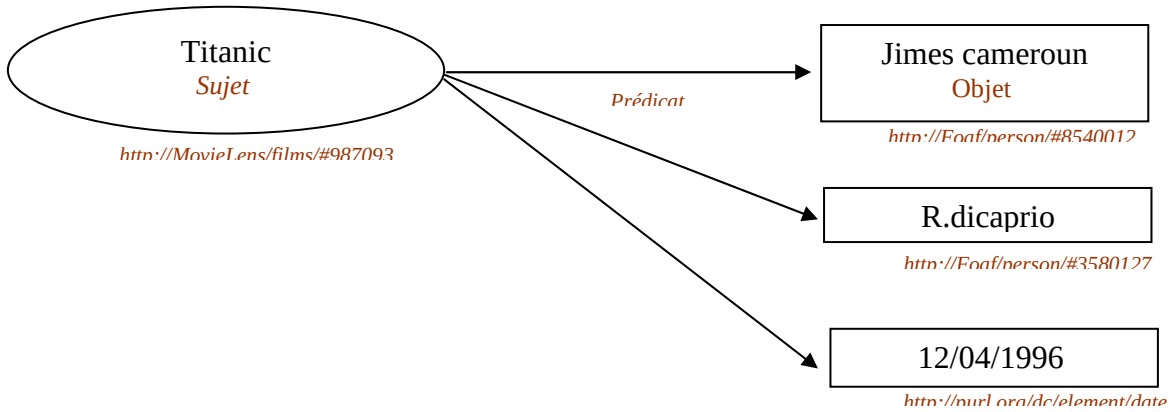
Les méta-données sont par définition des données concernant les données et dans le contexte du Web elles font référence à une information descriptive de ses ressources. Le Web a été initialement constitué pour un traitement humain et pour cette raison qu'il est difficile de tout y automatiser. Le concept de méta-données existait avant l'avènement d'Internet mais son intérêt a grandi avec le nombre de publications électroniques et de bibliothèques virtuelles. La solution proposée par le World Wide Web Consortium (W3C) est d'utiliser les méta-données pour décrire les données disponibles sur le Web. Dans le contexte du Web Sémantique elles constituent un module fondamental et permettent notamment de faciliter la recherche d'information. Elles garantissent l'interopérabilité en assurant le partage et l'échange d'information rendant son contenu lisible et compréhensible par les machines [GUI05].

Exemple

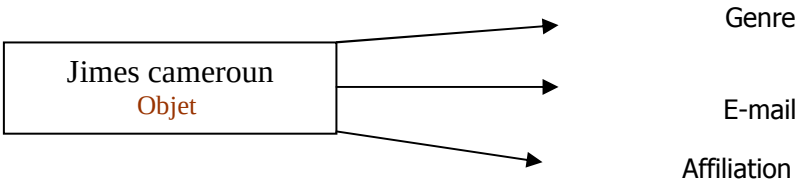
Le film " Titanic" á pour créateur "Jimes Cameroun".



Une ressource peut posséder plusieurs propriétés :



Du même un objet (utilisateur, acteur…) peut avoir plusieurs propriétés



6.2 Les annotations

Une annotation est une information graphique ou textuelle attachée à une ressource. Cette place est donnée par une ancre.
Les annotations peuvent prendre plusieurs formes comme :
. Des icônes (pour décrire des avis en utilisant des étoiles, des points d'interrogation...),
. Des symboles de liens (pour décrire des associations, des relations entre mots...),
. Des notes textuelles en marge, en bas de page ou en fin de document repérées dans le texte par des icônes (numéros, étoiles...),
. Des mises en formes typographiques (surlignage, soulignage, italique...),
. Des redécoupages de texte (à l'aide d'accolades, de numérotation de passages...),
. Des images, des sons, des concepts et leurs attributs (annotations sémantiques)...

6.3 Les systèmes d'annotations libres sur le Web

Les outils d'annotation libre doivent prendre en compte un certain nombre de contraintes et particularités du Web : les acteurs (lecteurs et serveurs Web) sont répartis, les communications se font par le réseau, le système est fondamentalement multi-utilisateurs (utilisateurs par ailleurs très nombreux), le langage de communication (HTTP), les données au format HTML ou XML... Globalement, tous les systèmes respectent le même schéma d'architecture (figure III.1) [VAS99] : un intermédiaire "observe" les transactions entre le client Web et les serveurs Web. Cet intermédiaire agit sur la requête, les pages obtenues et, éventuellement, sur les événements issus du navigateur. Cet élément est composé d'un intercepteur qui est chargé de récupérer requête et/ou pages HTML, d'un composeur qui se charge d'associer aux pages les annotations attachées (présentes dans une base de données). Cette combinaison peut dépendre du profil de l'utilisateur.
En pratique, les systèmes d'annotations libres sur le Web se divisent en deux grandes catégories : ceux basés sur des serveurs mandataires ("proxy") et ceux utilisant un intermédiaire attaché au navigateur (intermédiaire client).

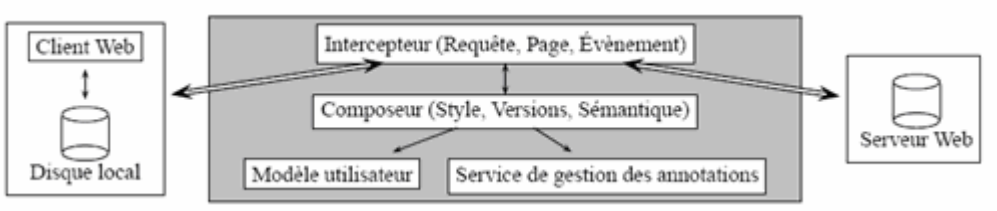


Fig III.1- Schéma d'un outil d'annotation sur le web

Amaya est un navigateur "open source" développé par le W3C, il intègre une application d'annotation collaborative (*Annotea Project*)[9] qui repose sur RDF et permet aux lecteurs d'une page de rédiger leurs propres annotations.

OntoAnnotate inclut des outils pour une annotation manuelle ou semi-automatique des pages [STA01].
SHOE [HEF00] fût l'un des premiers systèmes à proposer une annotation sémantique des pages Web. Il permet aux utilisateurs d'annoter les pages via des ontologies accessibles par des URIs ou localement ainsi que faire des raisonnements sur ces annotations.
Les travaux de [LER01] tentent d'annoter les pages automatiquement par des algorithmes d'apprentissage mais restent tributaires d'une longue phase d'apprentissage.
Dill *et al.* [2003] proposent *SemTag and Seeker*, un outil d'annotation automatique du Web en se fondant sur une base de connaissance ATP (avec une taxonomie lexicalisée des entités tels que le sport, la santé, etc.) ainsi qu'un nouvel algorithme de désambiguïsation.
Ils affichent d'excellents résultats en annotant 264 millions de pages Web, 434 millions annotations ont été générés.

7. Technologies et Standards W3C

Le consortium W3C[10] est une institution de recherche qui a comme mission la conduite de l'évolution du Web en préservant son interopérabilité, l'organisation mis en œuvre plusieurs standards et normalisations pour l'infrastructure web sémantique.

. FOAF

FOAF[11] *(Friend of A Friend)* est aussi un modèle de l'utilisateur en RDF. Il a été développé pour les communautés et groupes sociaux. L'utilisateur crée son fichier FOAF avec les informations de son choix :

. État civil
. Entreprise
. Liens avec d'autres personnes, nature de ces liens
. Projets
. Etudes

Des moteurs utilisant les données FOAF peuvent alors répondre à des questions telles que :
« *Quelles sont les personnes de moins de 30 ans habitant à Nantes, ayant un site web et aimant jouer au billard ?* ».

---

[9] http://www.w3.org/2001/Annotea

[10] http://www.w3c.org.

[11] http://www.foaf-project.org

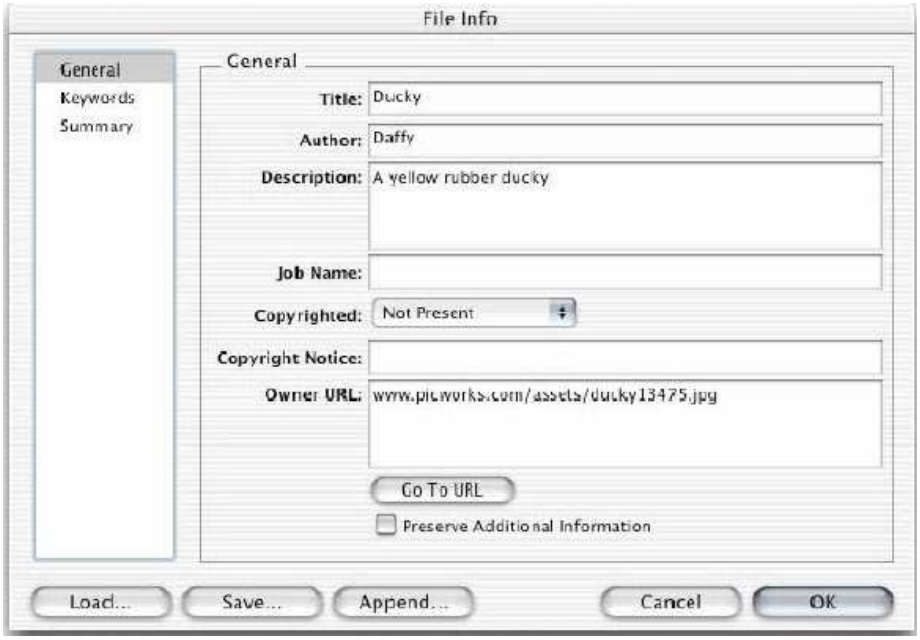


Fig III.2 -Fiche de saisie d'une ressource

. TAP

TAP [12] Projet mis de l'avant par
– Knowledge Systems Laboratory (Stanford)
– Knowledge Management Group (IBM Almaden)
– W3C
Propose un protocole GetData, pour récupérer sur le web des informations en format RDF Ce protocole permet d'identifier une ressource par ses propriétés plutôt que son URI.
Deux serveurs peuvent utiliser les descriptions pour identifier une même ressource qu'ils nomment de manière différente.
Cela suppose le partage d'un vocabulaire pour les descriptions d'une requête selon le protocole GetData qui contient les deux éléments suivants:
– Une description d'une ressource
– Une URI désignant une propriété
On s'attend à ce qu'un serveur recevant une telle requête retourne la valeur associée à cette propriété pour la ressource en question.
La requête est formulée dans le langage SOAP (langage défini pour l'accès aux services web).

TAP – exemple de requête GetData

```
<?xml version="1.0" encoding="UTF-8"?>
<SOAP-ENV:Envelope
xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
SOAP-ENV:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/"
xmlns:xsi="http://www.w3.org/1999/XMLSchema-instance"
xmlns:xsd="http://www.w3.org/1999/XMLSchema"
xmlns:a="http://tap.stanford.edu/">
<SOAP-ENV:Body>
<GetData>
<a:Musician>
<a:title>Yo Yo Ma</a:title>
<a:title>Ma, Yo-Yo</a:title>
<a:playsInstrument
```

[12] http://tap.stanford.edu/

```
ressource="http://tap.stanford.edu/kb/CelloInstrument"/>
<a:oid namespace="a">MusicianMa,_Yo-Yo</a:oid>
</a:Musician>
<a:concertSchedule/>
</GetData>
</SOAP-ENV:Body>
</SOAP-ENV:Envelope>
```

TAP a créé une base de connaissances contenant diverses données (musiciens, films, compagnies, etc.), quand un usager soumet un terme à l'outil de recherche, ce terme est recherché dans la base de connaissances, si on le trouve on oriente la recherche selon le type du concept et ses propriétés.

. Le Dublin Core

Le « ***Dublin Core*** » lancé en 1995 par l'OCLC (Online Computer Library Center) et le NCSA (National Center for Supercomputing Applications) comprend 15 éléments détaillés dans le tableau ci dessous.

Les promoteurs de ce projet sont partis d'un double constat :

- Le nombre de ressources internet disponibles augmente tous les jours, ce qui rend de plus en plus difficile la recherche d'informations précises.
- Les techniques de description actuellement mises en oeuvre sur internet ne permettent pas de répondre efficacement à cette augmentation du nombre de documents potentiellement accessibles.

Le Dublin Core2 [DCM03] est un vocabulaire (ou une ontologie) minimal pour l'indexation des pages web. Il a été défini sous l'égide de "Online Computer library Center", et maintenant d'un forum ouvert.

Tableau 1: *Les éléments du Dublin Core*

	Nom	Signification
1	TITLE	Nom donné à la ressource par l'entité CREATOR ou PUBLISHER
2	AUTHOR	Personne responsable du contenu intellectuel du DLO
3	SUBJECT	Description du sujet ou du thème dont traite la ressource
4	DESCRIPTION	Description en texte libre du contenu de la ressource
5	PUBLISHER	Personne ou institution en charge de la diffusion du DLO
6	CONTRIBUTORS	Personne ayant apporté une contribution intellectuelle à la ressource(en plus de l'entité désignée par AUTHOR)
7	DATE	Date de publication
8	TYPE	Le genre (au sens littéraire) auquel se rattache le DLO
9	FORMAT	Format du DLO
10	IDENTIFIER	Chaîne ou nombre utilisé pour identifier le DLO
11	SOURCE	Documents (sous forme imprimée ou électronique) dont ce DLO est dérivé
12	LANGUAGE	Langue dans laquelle est exprimé le contenu intellectuel de la ressource
13	RELATION	Relations avec les autres ressources
14	COVERAGE	Emplacement physique et les caractéristiques de durée de l'objet
15	RIGHTS	Gestion des droits

. Creative commons

Une initiative pour proposer aux créateurs de contenu qui veulent le distribuer de manière ouverte, un moyen de produire très rapidement une licence d'utilisation. L'intérêt de cette licence est qu'elle est simultanément engendrée en trois formats à partir d'une source unique, une version à l'aide de pictogrammes pour les gens normaux, une version en RDF destinée à être traitée automatiquement et une version en langage juridique pour les avocats.

<rdf:RDF xmlns="http://web.resource.org/cc/"

xmlns:dc="http://purl.org/dc/elements/1.1/"

xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">

<Work rdf:about="">

<dc:type rdf:resource="http://purl.org/dc/dcmitype/Text" />

<license rdf:resource="http://creativecommons.org/licenses/by-ncnd/

2.0/" />

```
</Work>
<License rdf:about="http://creativecommons.org/licenses/by-nc-nd/2.0/">
<permits rdf:resource="http://web.resource.org/cc/Reproduction" />
<permits rdf:resource="http://web.resource.org/cc/Distribution" />
<requires rdf:resource="http://web.resource.org/cc/Notice" />
<requires rdf:resource="http://web.resource.org/cc/Attribution" />
<prohibits rdf:resource="http://web.resource.org/cc/CommercialUse" />
</License>
</rdf:RDF>
```

. VCARD

(VisitCard)[13] définit une représentation informatique pour l'échange des données personnelles (Nom, prénom, téléphone, e-mail, date de naissance, URL, etc.), il est utilisé par les logiciels de carnet d'adresses (Outlook, Thunderbird, …) ainsi que par les appareils mobiles et les logiciels de messagerie.

. Microformats

Les micro-formats, sont des attributs de classe particuliers, appartenant à une liste prédéfinie par une communauté. Ils sont ajoutés aux balises du code HTML, et jouent alors un double rôle de présentation de style et de structuration sémantique.

Exemple

Pour inscrire de la page d'affichage du navigateur : http://Benameur.dz/
Le code HTML non microformaté

```
<div>
<a href = http://benameur.dz>
 <span> Benameur</span>
 <span>ali </span>
</a>
</div>
```

Code html avec implémentation du microformat hCard

```
<div class= "vcard">
<a classe="url" href="http://beameur.dz/">
<span class="given-name"> Benameur </span>
<span class="family-name"> ali </span>
</a>
</div>
```

[13] (ftp://ftp.isi.edu/in-notes/rfc2426.txt)

L'objectif général du projet Microformats est de permettre le passage d'un Web d'affichage des contenus par le biais du navigateur, à un écosystème sémantique avec lequel il est possible d'interagir pour extraire les informations utiles afin de les exploiter ailleurs.
En effet, non seulement les microformats permettent-ils de préciser et structurer les informations contenues dans une page Web, mais surtout ils facilitent la réutilisation de ces informations structurées par un ensemble de logiciels compatibles microformats. Ces données pourront ainsi être détectées, identifiées et exploitées par des applications multiples (moteurs de recherche, agrégateurs, applications de bureau, gestionnaires de contacts et des utilitaires de calendriers, etc.).

8. Indexations sémantique

L'indexation sémantique *(Sense Based Indexing)* est une approche d'indexation basée sur le sens des mots [SAN94]. Elle s'appuie sur des algorithmes de désambiguïsation de mots (WSD) pour indexer les documents et les requêtes avec le sens des mots (mots-sens) plutôt qu'avec des mots simples. Une manière d'indexer serait par exemple, d'associer aux mots extraits, des mots du contexte qui aident à déterminer leur sens : par exemple, *bank(river/money*) et *plant(manufacturing/life)*. D'autres approches de désambiguïsation plus élaborées, utilisent des représentations hiérarchiques pour calculer la *distance sémantique* ou *similarité sémantique* entre les mots à comparer [LEK94]. Cette notion de distance sémantique est plus générale et peut aussi être utilisée
Nous distinguons deux types de démarches dans l'indexation sémantique : la démarche issue de la RI et la démarche issue du Web Sémantique.

8.1 La démarche issue du domaine de la RI

Consiste à choisir comme langage de représentation des documents, l'ensemble des concepts et instances de l'ontologie (hiérarchies de concepts). Les descripteurs ne sont plus choisis directement dans les documents ou dans un vocabulaire contrôlé (ou thésaurus) mais au sein même de l'ontologie. Les granules documentaires sont alors indexés par des concepts qui reflètent leur sens plutôt que par des mots bien souvent ambigus [AUS04]. Il convient dans ce cas d'utiliser une ontologie reflétant le ou les domaines de connaissance abordés dans la collection documentaire. Il est en effet nécessaire de retrouver dans l'ontologie les concepts présents dans la collection pour indexer les documents à partir de toutes les thématiques abordées.
Les ontologies de domaine peuvent par leur formalisation représenter des ressources impliquant.

8.2 La démarche orientée Web Sémantique

Les précurseurs de cette nouvelle version du Web considèrent que les ressources participant au WS seront toutes reliées entre elles par des relations sémantiques. Plus précisément, les données présentes sur le WS seront modélisées sous forme d'ontologies où chaque ressource apparaît comme un élément de ces ontologies au même titre que la connaissance qui les décrit. L'objectif est donc d'ajouter au contenu du Web une structure formelle et de la sémantique (à travers des méta-données et de la connaissance) dans le but de permettre une meilleure gestion et un meilleur accès aux informations. Cette démarche repose sur des ontologies modélisant les objets du monde à travers les acteurs et entités que les documents constituent et comportent [GUH03]. Elles peuvent être vues comme une représentation des méta-données explicitement ou implicitement présentes dans les documents. La phase d'indexation est aussi appelée annotation de documents. L'annotation de documents a pour but de représenter les informations relatives au média (date de création, taille, format d'encodage), les méta-données présentes dans les documents (auteurs, date de production), les index (les descripteurs du contenu du document), l'identifiant du document par le système (emplacement) et une vue sur le contenu (résumé ou extraits) [EUZ02]. La mise en place de

cette nouvelle vision du Web dépend de la présence de ces méta-données. Un enjeu actuel du WS est de définir des techniques permettant de les extraire [KIR04]. La démarche orientée Web Sémantique a donc un double objectif : indexer le contenu des documents à partir des ressources permettant d'en extraire les concepts et instances mais aussi représenter les ressources en générant les méta-données correspondantes.

9. Similarité sémantique

La similarité sémantique a intéressé diverses communautés de recherche en intelligence artificielle, psychologie et sciences cognitives. Elle a comme principaux champs d'application la recherche d'information et le traitement automatique de la langue.
Il est important de noter que dans la littérature, on parle aussi de proximité sémantique (*semantic relatedness*) qui est une notion plus large que la similarité sémantique. En effet la proximité sémantique prend en considération tout type de relation entre concepts. Ainsi deux concepts peuvent être proches sémantiquement par leur similarité (e.g. voiture et automobile), mais aussi par d'autres relations comme *partie-de* (voiture-roue) ou *contraire* (guerre-paix), etc [HAI05].

Deux approches principales sont utilisées pour la mesure de similarité entre concepts dans une ontologie : (i) en utilisant la structure arborescente ou (ii) en utilisant le contenu informatif des différents concepts en intégrant des mesures statistiques. D'autres approches proposent de combiner les deux, la plupart de ces propositions portent sur WordNet elles peuvent cependant être généralisées à une ontologie puisqu'elles exploitent la structure taxonomique.

On définit formellement la Similarité par la fonction :

$S2 \rightarrow [0; 1]$ avec S l'ensemble de concepts.

. Caractéristiques

D'après [LIN98] la similarité entre deux concepts A et B respecte :
Fonction des caractéristiques communes. Plus les entités ont des caractéristiques en commun, plus elles sont similaires.
Fonction de leurs différences. La similarité décroît inversement aux caractéristiques différentes.
Maximale quand A est identique à B.

Propriétés:

sim(x,x)=1 : la réflexivité
sim(x, y)=sim(y, x): la symétrie

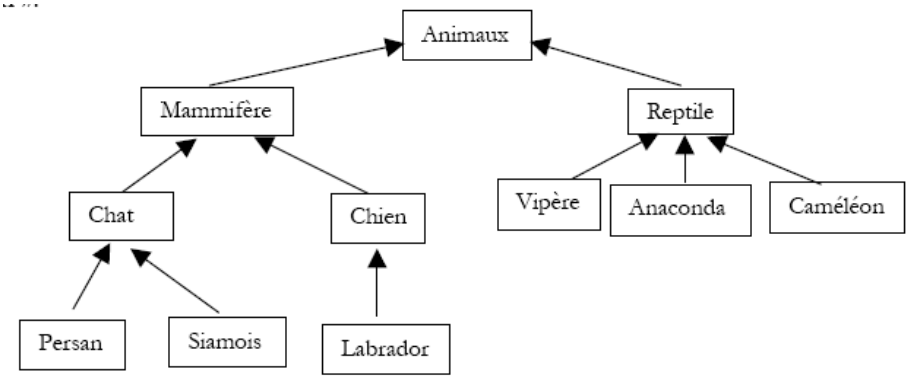
Remarque :
La similarité est la fonction inverse de la distance, plus deux mots sont similaires, moins ils sont distants de façon que $dist(a; a) = 0$, $dist(a; b) = dist(b; a)$ et $dist(a; b) \leq dist(a; c) + dist(c; b)$.

9.1 Calcul de similarité par distance

Les mesures reposant sur la distance considèrent que la similarité entre deux concepts peut être calculée à partir du nombre de liens qui séparent les deux concepts. Plusieurs variantes existent en fonction du chemin pris en compte pour calculer la distance entre les concepts.

Exemple

La figure II.8 est utilisée pour illustrer les différentes mesures. Les concepts sont représentés par des rectangles et les flèches symbolisent la relation « est un » [HER05].



FigII.8 -Exemple de taxonomie

. Mesure de RADA

Rada propose La mesure du **edge counting** qui évalue la distance sémantique à partir du nombre de branches séparant les concepts par le plus court chemin dans la hiérarchie [RAD89].

D'après l'exemple ci-dessus

$Dist_{edge}(siamois, persan)=2$

$Dist_{edge}(reptile, anaconda)=1$

$Dist_{edge}(chat, mammifère)=1$

$Dist_{edge}(persan, reptile)=4$

$Dist_{edge}(persan, labrador)=4$

La limite d'utiliser le chemin le plus court est qu'on ne prend pas en considération la position des concepts dans l'ontologie. Intuitivement, deux concepts classés en bas de l'ontologie sont très spécifiques et sont donc à un degré de granularité plus fin que deux concepts classés en haut de l'ontologie.

. Mesure de leacock

Leacock a proposé la formule suivante pour calculer la similarité. Elle est issue de la proposition de Resnik et assure la normalisation de cette dernière [LEA98].

$$\mathrm{Sim}_{\mathrm{edge}}(c1,c2) = -\log \frac{Dist_{edge}(c1,c2)}{2*Max}$$

Où max étant la profondeur maximale de la taxonomie.
L'utilisation de la fonction –log permet de normaliser la similarité entre [0,1] (1 signifiant que les concepts sont totalement similaires).

$sim_{edge}(siamois, persan) = -log\ 4*22=0,6$

sim_{edge} *(reptile, anaconda)= 0,9*
sim_{edge} *(chat, mammifère)=0,9*
sim_{edge} *(persan, reptile)=0,3*
sim_{edge} *(persan, labrador)=0,3*

. Msure Wu et Palmer

Wu et Palmer ont proposé une autre mesure de similarité prenant en compte à la fois la profondeur des concepts dans la hiérarchie de concepts et la structure de la hiérarchie de concepts.
Pour calculer la similarité entre deux concepts c1 et c2, la formule suivante est utilisée :

$$\mathrm{Sim}_{\mathrm{Wu}}(c1,c2)=\frac{2*depth(c)}{depth(c1)+depth(c2)}$$

où depth(c_i) correspond au niveau de profondeur du concept c_i et c représente le concept le plus spécifique qui généralise c1 et c2.
La valeur de la similarité est comprise entre 0 et 1 (1 signifiant que les concepts sont totalement similaires, 0 totalement différent).

sim_{Wu} *(siamois, persan)= 2*3/(4+4)=3/4 (chat étant le plus spécifique subsumeur)*
sim_{Wu} *(reptile, anaconda)= 2*2/(2+3)=4/5* sim_{Wu} *(chat, mammifère)=4/5*
sim_{Wu} *(persan, reptile)=2*1/(2+4)=1/3* sim_{Wu} *(persan, labrador)=2*2/(4+4)=1/2*
Cette mesure est plus pertinente que les mesures précédentes reposant uniquement sur le chemin le plus court entre les deux concepts, car elle prend en compte l'organisation hiérarchique des concepts, c'est-à-dire le concept généralisant les deux concepts considérés.

9.2 Calcul de similarité par le contenu informatif

La notion de contenu informatif (CI) a été introduite pour la première fois par Resnik. Elle utilise conjointement l'ontologie et un corpus. Le contenu informatif d'un concept traduit la pertinence d'un concept dans le corpus en tenant compte de la fréquence de l'apparition des mots auxquels il se réfère ainsi que de la fréquence d'apparition des concepts qu'il généralise. Plus précisément le contenu informatif se calcule par la formule suivante :

$$CI(C) = -log(p(C))$$

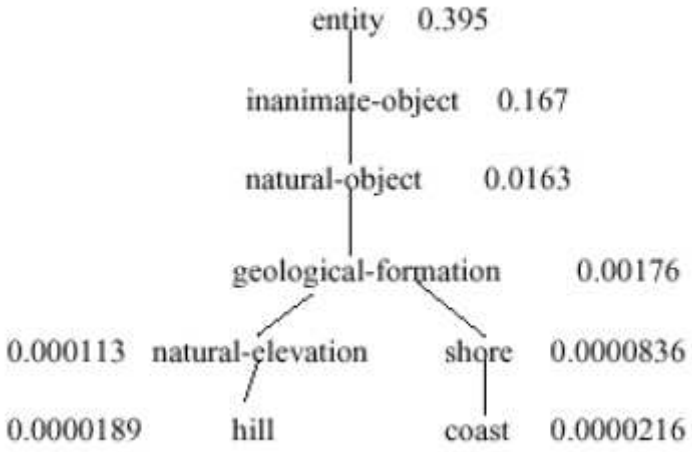
Où $p(C)$ est la probabilité de retrouver qu'un mot du corpus soit une instance du concept C (un des mots référés par le concept C ou par un de ses descendants), elle est monotone quand on remonte dans la hiérarchie ($p(A) \leq p(B)$) si A est plus grand B.
Dans les expérimentations de [RES95], ces probabilités sont calculées par

$$p(C) = \text{frequence}(C)/N$$

N est le nombre total de concepts et $frequence(C) = \sum_{w \in instance(C)} count(w)$.

Plus un concept est général, plus son contenu informatif est faible, ainsi $CI(p(T)) = 0$ parce qu'il a un contenu informatif nul. A l'inverse, plus le concept est spécifique plus son contenu

informatif est grand. L'intuition de la notion de contenu informatif est que la similarité entre deux concepts est la portion d'information qu'ils ont en commun qui, dans le cadre d'une ontologie, peut être déterminée par le concept le plus spécifique qui les subsume (*ppac*). Cette intuition est indirectement appliquée par les mesures présentées dans la section précédente qui calculent la similarité avec le nombre d'arcs qui séparent deux concepts.



FigureII.9 Extrait de WordNet [Lin98]
(*Avec les probabilités correspondantes aux différents concepts*).

Resnik définit la similarité sémantique entre deux concepts par la quantité d'information qu'ils partagent. Cette information partagée est évaluée numériquement par le contenu informatif du plus petit ancêtre commun (*ppac*).

$$simRes(C1;C2) = CI(ppac(C1;C2))$$

Ainsi, si deux termes sont très éloignés et ont comme *ppac* la racine, leur similarité est égale à 0. L'approche de Resnik essaie d'éviter le problème de granularité, cité au dessus, en diminuant le rôle des arcs dans le calcul de similarité.

9.3 Calcul de similarité par hybridation

Jiang et Conrath proposent en combinant le contenu informatif du *ppac* à ceux des concepts. Elle prend en considération aussi le nombre d'arcs en calculant le contenu informatif de chaque concept [JIA97].
Ainsi une distance est définie :

$$distance(C1;C2) = CI(C1) + CI(C2) - (2{:}CI(ppac(C1;C2)))$$

La mesure de similarité revient donc à calculer :

$$sim_{JC}(C1;C2) = 1/\ distance(C1;C2)$$

Lin, propose une mesure de similarité qui calcule la proportion d'information commune entre deux concepts par rapport à leur description[LIN98].

$$sim_{Lin}(C1;C2) = 2 * CI(ppac(C1;C2))/CI(C1) + CI(C2)$$

10. Méthodologies
10.1 Techniques Collaboratives

- L'évaluation des ressources par les utilisateurs

Le cas le plus simple est de demander á un nouveau utilisateur de noter un ensemble de ressources, comme un échantillon de référence selon une échelle bien déterminée (surmonter la phase de démarrage á froid).

- Stockage des données

Typiquement, on représente ces données par une matrice de notes des utilisateurs sur les ressources qu'ils ont parcourues.

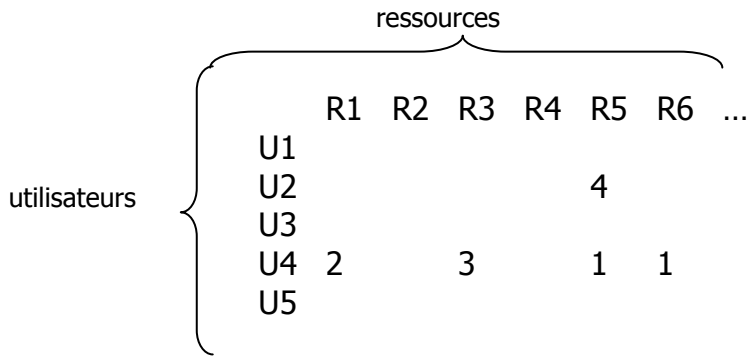


Fig II.10- Matrice d'évaluation

Cette matrice correspond á un ensemble de votes $v_{i,j}$. Par exemple ici, on a $v_{2,5}$=4 c'est-á-dire que le vote de l'utilisateur 2 pour la ressource 5 est égale á 4.
$\vec{v_i}$ correspond au vecteur décrivant l'utilisateur i, $\vec{v_4}$ =(2,0,3,0,1,1).

- Calcul de prédiction

La méthode de l'exploitation des données stockées pour aider les utilisateurs dans leurs recherches (la prédiction), est une étape cruciale dans un système de recommandation.
Selon [BHK98] il y a deux grandes classes de méthodes pour calculer la prédiction :
Les *méthodes basées sur la mémoire* utilisent l'entièreté de la base de données utilisateur pour faire des prédictions et les *méthodes basées sur les modèles* utilisent la base de données utilisateur pour estimer ou apprendre un modèle, qui sera ensuite utilisé pour les prédictions.

- Mise à jour du profil utilisateur

A partir des jugements de l'utilisateur le système prend en charges ces actions et rafraîchir la base de donnée de l'utilisateur (*profil*) afin d'une meilleure prédiction ultérieure.

Or, ces techniques montrent quelques difficultés :

- Ces méthodes sont basées essentiellement sur la classification et les statistiques ce qui conduit á une complexité combinatoire.
- Le démarrage á froid et le passage á l'échelle.
- Peu ou pas des utilisateurs ont évalués des ressources (matrice creuse).
- L'insuffisance en nombre pour la comparaison par mesure de similarité.

Pour faire face á ces situations critiques ont fait recoure aux d'autres techniques pour améliorer le filtrage d'information.

10.2 Techniques Sémantiques

Dans notre contexte, la notion de similarité est plutôt celle de la similarité sémantique, qui est également appelée la proximité sémantique. Elle est déterminée grâce à l'association à des ressources ou des entités, d'une métrique basée sur la similitude de leurs significations ou de leurs contenus sémantiques. La similarité est la quantité qui reflète la force du rapport entre deux objets ou deux caractéristiques.

Concrètement, ceci peut être réalisé par exemple en définissant une similarité topologique, ou en employant des ontologies pour définir une distance entre les objets (ou les noeuds représentent les utilisateurs ou les ressources et les arcs la distance).

La distance mesure la dissimilarité de deux entités, elle est inverse de la similarité : si la valeur de la fonction de similarité de deux entités est élevée, la distance entre elles est petite et vice-versa [EUZ04].

Les centres d'intérêts de l'utilisateur sont appariés aux concepts des domaines de l'ontologie. A l'aide de mesure de similarité (par distance ou contenu informatif), on regroupe les utilisateurs les plus similaires selon diverses propriétés :

- même centre d'intérêt.
- même génération (age).
- Zone géographique.
- Activités sociales
- Loisirs ….

De la même façon on associes les ressources entre eux et on fait une projection sur l'ontologie du domaine tout en déduisant une similarité selon diverses caractéristiques : (dans notre cas typique on substitue une ressource par un film extrait du champs de test MovieLens)

- Titre de film
- Acteurs
- Date de sortie
- Catégorie
- Durée

10.3 Techniques Hybrides

Notre Objectif est la combinaison entre l'approche collaborative est l'approche sémantique pour limiter les inconvénients tels que le démarrage á froid, matrice creuse et le facteur de pertinence. En effet, la synergie de données sémantique permet de booster le processus de filtrage et augmente ces fonctionnalités.

L'hybridation des méthodes peut être effectuée avec différentes méthodes, décrites dans [BUR02] parmi lesquelles : la pondération, le changement adaptatif, la composition, la combinaison, la cascade, l'augmentation de fonctionnalité, etc.

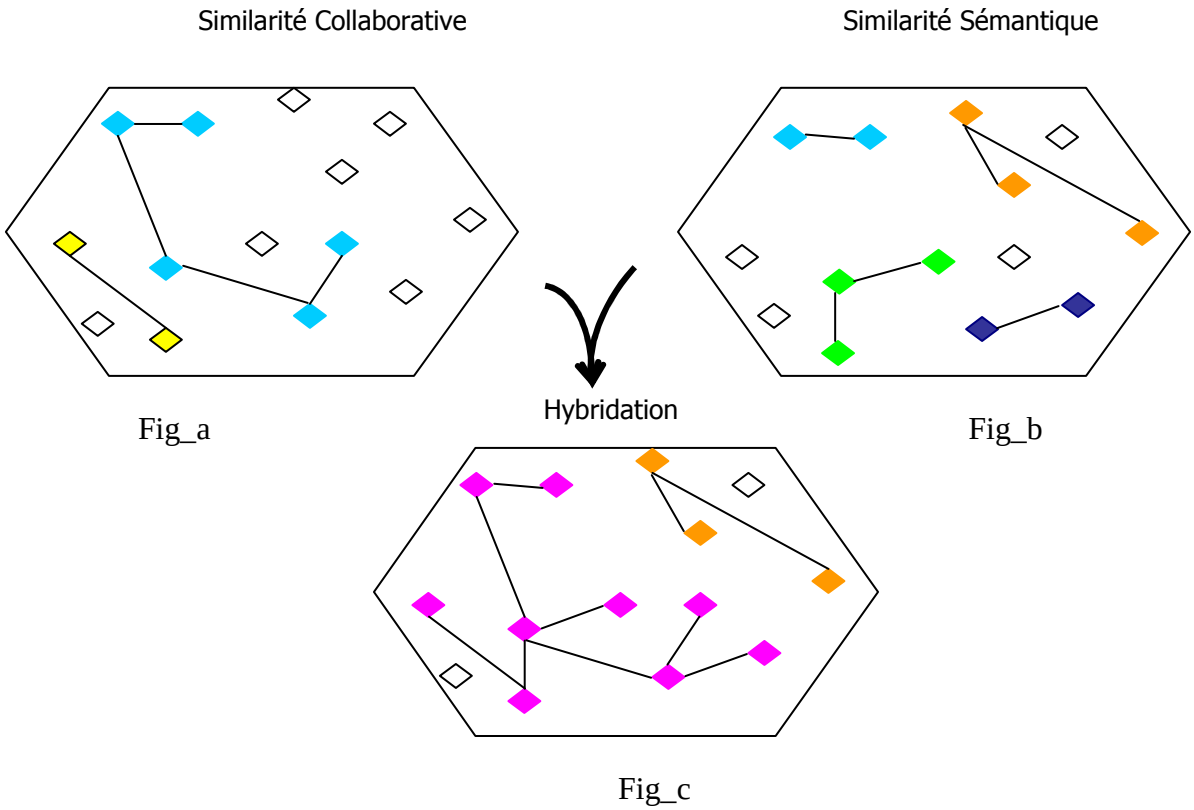


Fig - II.11 Combinaison de similarité Sémantique/Collaborative

◇ Ressource
—— Similarité (Collaborative, fig_a)
(Sémantique, fig_b)
(Collaborative+sémantique, fig_c).

On note bien, que l'approche sémantique illustré dans la fig_b vient de compléter les liens entre les ressources décrites dans la fig_a ce qui augmente le facteur de similarité et le regroupement des ces ressources fig_c.

11. Apport du web sémantique pour le filtrage d'information

Le web sémantique et l'utilisation des ontologies offrent des avancées majeures :

- L'augmentation de la pertinence via la description des ressources par les méta-données.
- Recommandations améliorées, elles prennent en compte les liens unissant les entités de même portent les sens et contexte d'utilisation.
- De plus, chaque élément des documents étant représenté á l'aide d'une ontologie, il devient très facile d'automatiser des tâches définies á partir des méta-données.
- L'interopérabilité et l'automatisation des tâches entre les entités dans le système permet d'atténuer l'ardu de visualisation par l'utilisateur ainsi un gain de temps considérable.
- En effet le web sémantique montre une utilité remarquable et un degré de confiance important pour les systèmes de recommandation, ce qui encourage beaucoup les

utilisateurs de s'orienter vers cette recherche passive qui satisfait largement leurs besoins en information par rapport aux méthodes classiques.

12. Difficultés de l'approche WS

D'une part, les outils permettant de créer et de gérer les ontologies ne sont pas encore assez matures pour permettre " l'industrialisation " de l'utilisation des ontologies dans les entreprises. Ils ne sont pas d'accès faciles et il n'existe pas encore de véritables standards. Toutefois, certains éditeurs de logiciels, comme Mondeca, proposent des solutions complètent qui intègrent les différents standards.
D'autre part, il n'existe pas de méthodologies prouvées et surtout éprouvées pour guider les entreprises dans la création d'ontologies. La création d'une ontologie est une affaire de spécialiste dans la mesure où il faut avoir des compétences à la fois dans le domaine à modéliser mais aussi en représentation des connaissances et linguistiques.
Il parait difficile d'accompagner chaque ressource par ces propres méta-données structurées de manière fiable et standardisée.
Gageons que les organisations qui entreprennent dès aujourd'hui la mise en place de Serveurs de connaissances basés sur l'infrastructure du web sémantique possèderont une avance certaine quant à l'exploitation de leurs ressources immatérielles[GIL08].

Conclusion

Nous avons présenté dans ce chapitre, l'architecture de base en couches du web sémantique, sa propagation tient á l'acceptation des standards et la facilitation de la mise en œuvre. L'utilisation de telle ontologie s'avérer nécessaire pour mieux cibler l'information et extraire le concept visé par projection ou par une autre mesure. L'enrichissement des systèmes de filtrages d'information par le biais de cette infrastructure réside dans le passage de l'aspect classique basé essentiellement sur le traitement statique d'un ensemble de mots et parfois d'un balayage total des documents vers la structurations et l'annotation des documents ainsi qu'une modélisations sophistiqué des profils par l'adjonction de nouveaux dimensions notamment sémantique qui permet d'augmenter l'effet de pertinence et donne une nouvelle construction de groupes utilisateurs fortement liées pour une meilleure recommandation attendue par ces systèmes.

Partie II

Proposition et validation

CHAPITRE III

OBJECTIFS ET PROPOSITION

1. Objectifs

La quantité d'informations augmente beaucoup plus rapidement que notre capacité à traiter. Il est temps maintenant de créer les technologies qui peuvent nous aider à passer par toutes les informations disponibles afin de découvrir celle qui est plus précieuse pour nous. L'une des plus prometteuses de ces technologies est le filtrage collaboratif (FC).
Le processus de FC est le noyau des systèmes de recommandations qui ont été largement utilisés dans le commerce électronique tels que Amazon.com, CDNOW, MovieLens,…etc. En réalité, la plupart des systèmes de recommandation modernes face à des dizaines de milliers (voir des millions) d'utilisateurs et d'objets.
L'objectif de notre étude est d'apporter une amélioration par différents points de vus pour ces systèmes qui marquent quelques défis, parmi lesquels :

a- Évolutivité de la mémoire (Scalability)

Appelé aussi passage á l'échelle cela dû quand l'algorithme traite une grande quantité d'information (millions d'items, millions d'utilisateurs) pour trouver les K plus proches voisins pour l'utilisateur actif, le problème s'aggrave quand les calculs se font en ligne (mode on-line), par exemple dans un site web visité par un grand nombre de personnes sur des milliers d'items cela provoque une difficulté pour la formulation d'une recommandation en temps réel.

b- L'indisponibilité de données d'évaluation (Sparsity)

Appelé aussi matrice creuse, suite au nombre très élevé d'items, l'évaluation des utilisateurs est très peu. Prenons un exemple d'un système de recommandation des livres sa base de données contient 2 millions de livres , l'évaluation est moins de 1% (200 000 livres) donc il est moins probable de trouver des évaluations communes entre utilisateurs (un minimum de voisins similaires) dans ce cas la comparaison de l'utilisateur actif (cible) avec ses proches est réduite, ce qui baisse directement la qualité de recommandation.

c- Le démarrage á froid : (could start)

Cela pour un nouveau utilisateur (ou item) intégré au système sachant que le calcul de la similarité dépend de l'historique de l'évaluation de l'utilisateur actif et celui des autres voisins similaires, dans notre cas son historique (profil) est vide, alors que le système fournit des recommandations qui peuvent être insatisfaisantes pour ce nouveau utilisateur [MEL02].

À certains égards, trouver un moins de temps consacré á un algorithme de recherche voisins et augmenter la qualité de recommandation semble á une manipulation d'une balance (deux objectifs en conflit). Pour cette raison, il est important de traiter les deux objectifs simultanément afin de découvrir les solutions à la fois utile et pratique.

2. Proposition

Dans cette partie nous présentons les points suivants comme réponses aux défis cités ci-dessus :

Premièrement l'adoption de l'approche basée items pour l'algorithme de filtrage collaboratif á l'inverse des systèmes classiques qui sont basés sur l'approche basés user (mode on-line) pour le calcul de la similarité.
Cette approche suggérée est caractérisée par :

- Le calcul de la similarité entre items se fait en mode off-line (temps régulier, mode batch,…) ce qui augmente les performances du système.
- L'espace des items est relativement faible par rapport á l'espace des utilisateurs ce qui raccourci le calcul des corrélations entre ces items.

Deuxièmement la possibilité d'intégrer des données extérieures (données démographiques, méta-données, informations sémantiques, sens des relations entre items,…) au système pour réduire le problème de l'indisponibilité de l'évaluation et réduire l'effet du démarrage á froid par l'exploitation d'une ontologie de domaine.

Finalement, nous avons appliqué une technique LSI (Latent Semantic Indexing), pour réduire l'espace de traitement initial en gardant au maximum l'importance de l'information (décomposition en valeurs singulières SVD).
Le module collaboratif basé item est détaillé dans la section 2.1, le module sémantique est détaillé dans la sections 2.2 et la technique LSI est décrite dans la section 3.

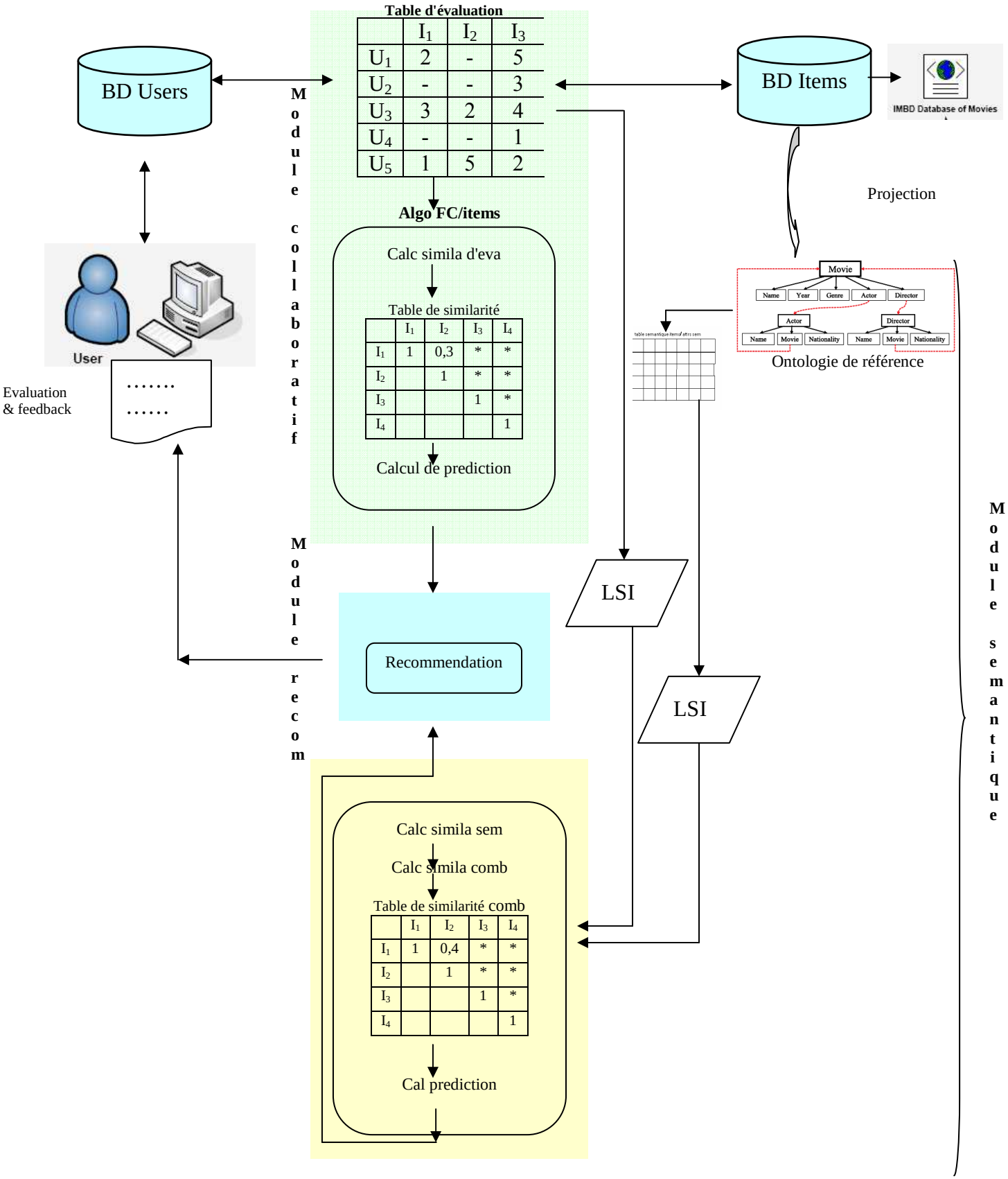


Fig III.1- Architecture globale du système

2.1 Approche basée items

L'intuition derrière cette approche est que l'utilisateur est intéressé à des objets qui sont semblables, cela mène á étudier les relations entre items et les classifier plutôt que chercher la similarité entre un nombre élevé d'utilisateurs (les SFC classiques), la tendance de ces systèmes est de produire beaucoup plus rapidement des recommandations.

- Calcul de la similarité des items

Pour calculer la similarité entre l'item I_p et I_q on doit d'abord identifier tous les utilisateurs qui ont évalué les items I_p et I_q (deux vecteurs), puis on utilise une mesure pour calculer la similarité entre ces deux vecteurs par exemple la mesure de cosinus :

$$\text{sim}(i_p,i_q)=\frac{i_p \bullet i_q}{\sqrt{\|i_p\| * \|i_q\|}} \qquad (1)$$

Où • est le produit scalaire.
Ou bien la mesure de corrélation :

$$\text{sim}(i_p,i_q)=\frac{\sum_{k=1}^{m}\left(R_{k,p}-\overline{R}_p\right).\left(R_{k,q}-\overline{R}_q\right)}{\sqrt{\sum_{k=1}^{m}\left(R_{k,p}-\overline{R}_p\right)^2.\left(R_{k,q}-\overline{R}_q\right)^2}} \qquad (2)$$

Où k=1..m : la liste des utilisateurs évaluant les items I_p et I_q.
$R_{k,p}$: la valeur de l'évaluation de l'utilisateur k pour l'item p.
$\overline{R}_p$: la moyenne de l'évaluation de l'item p.

Remarque : le résultat de ce processus est une matrice carrée de similarités entre items (diagonale supérieure ou inférieure).

- Calcul de la prédiction

Apres le calcul de la similarité, on sélectionne les items les plus similaires (les K plus proches voisins) pour l'item courant ensuite on génère la valeur de prédiction pour cet item à travers les évaluations de l'utilisateur courant pour les K items similaires.
La somme des poids nous donne:

$$R_{a,k}=\frac{\sum_{t=1}^{K}\left(R_{a,t}.sim(i_k,i_t)\right)}{\sum_{t=1}^{K}sim(i_k,i_t)} \qquad (3)$$

Où $R_{a,t}$: est la valeur de l'évaluation de l'utilisateur courant a sur t^{ieme} item similaire.
K: la taille des items les plus similaires.

- Génération de la liste de recommandation

Cette étape est automatique et on procède la génération de la liste de recommandation qui comporte les items ayants les valeurs de prédictions les plus élevées (N-top list), aussi un item est jugé utile (recommandé par le système) si sa valeur de prédiction est supérieure à un seuil donné.

2.2 La connaissance sémantique

Comme indiqué ci-dessus, le seul critère pour mesurer la similitude est la valeur de l'évaluation (par exemple note positive dans l'échelle 1..5), mais cette mesure dépend en réalité des autres dimensions telles que couleur, forme, age, poids, catégorie, centre d'intérêt …etc. Donc une modélisation par mesures implicites s'avère utile pour améliorer la précision des systèmes de filtrage.

- Les motivations de la connaissance sémantique

L'approche discutée en 2.1 sépare la tâche de calcul de similarité (hors ligne au lieu en temps réel) et la tâche de recommandation, ce qui permet d'ajouter l'information sémantique et par conséquent d'ajuster la classification des items afin d'une meilleure prédiction.

En outre, les attributs sémantiques des items donnent les raisons implicites d'un utilisateur d'être intéressé ou pas par tels items, cela à son tour, permet au système de faire des inférences sur la base de cette source supplémentaire de connaissance et améliore la recommandation.

Un autre avantage pour un nouveau item qui se heurte au problème " non évaluation", alors on fait recours aux informations sémantiques et les relations implicites pour la génération de prédictions via la similitude sémantique.

- L'ontologie de référence

Afin d'obtenir des informations sémantiques sur les éléments utilisés dans le processus de FC il faut extraire les entités sémantiques comme des objets structurés du domaine. Cette tâche implique l'extraction automatique et la classification des objets selon une ontologie du domaine, une ontologie est une représentation de connaissances au niveau conceptuel. Elle est toujours liée à un domaine particulier de connaissances qui est un ensemble borné et cohérent de connaissances doté d'une sémantique consensuelle [BBZ95]. Une ontologie contient les primitives terminologiques du domaine (le vocabulaire conceptuel) ainsi que des axiomes qui restreignent l'interprétation des primitives. Le vocabulaire conceptuel est, de manière consensuelle, structuré en un ensemble de concepts et un ensemble de relations existantes entre ces concepts.
Un exemple d'une ontologie comme un schéma relationnel d'une base de données comportant plusieurs tables liées par des clés sémantiques. Cette construction a pour but d'amplifier le niveau significatif de la connaissance.
Notre objectif est d'extraire les valeurs des attributs sémantique des objets de l'ontologie correspondantes á nos items afin de calculer les mesures de similarité.
Par exemple, certains utilisent des agents logiciels exploitants des heuristiques pour l'extraction automatique de l'information á partir des sites web.

A titre d'exemple considérons la base de données de site web MovieLens (www.imdb.com) constituée des pages décrivent les films (movies), une page peut inclure les attribut d'un film tels que le titre, le réalisateur, les acteurs ..etc, ceux-ci référencent les attributs associés à une classe qui représente les films dans l'ontologie de référence.

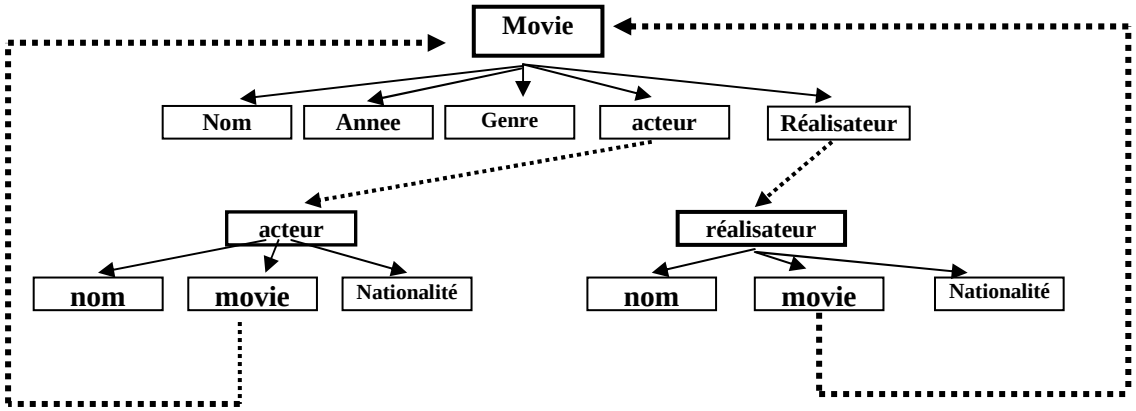


Fig III.2 - Portion de l'ontologie du web site (MovieLens)

Dans la figure certains attributs représentent les propriétés d'une classe donnée et les autres représentent des références correspondantes à d'autres classes.
La collection de pages Web dans le site représente un groupe d'objets incorporés qui sont les instances de ces classes, les attributs "Acteur" et de "réalisateur" référencent d'autres classes dans l'ontologie.
Dans notre cas, les valeurs des attributs sémantiques associées à ces classes sont rassemblées dans un tableau relationnel dont les lignes représentent les n items, et dont les colonnes correspondant à chacun des attributs extraits. On obtient une matrice de n items (lignes), et m attributs sémantiques appelée une matrice des attributs sémantiques.

Remarque : une tache de normalisation et de discrétisation est nécessaire pour la représentation uniforme des données (continues et symboliques) et les collectées dans des tables.

2.3 Combinaison de la similarité sémantique et l'évaluation collaborative

On combine les approches de calcul de la similarité entre items :

La similarité totale $SimTot(i_p,i_q) = \alpha\, SimSem(i_p,i_q) + (1-\alpha) SimEval(i_p,i_q)$. $0 \leq \alpha \leq 1$ *(4)*

Où $SimSem(i_p,i_q)$ est la similarité sémantique entre les items i_p et i_q calculée à partir des valeurs extraites de la matrice sémantique calculées par une technique de comptage d'arcs.
$SimEval(i_p,i_q)$ est la similarité d'évaluation par mesure de cosinus.
α paramètre ajusté selon le les résultats expérimentaux.

Si $\alpha=0$ l'approche est purement collaborative.
Si $\alpha=1$ l'approche est purement sémantique.

On procède de changer la valeur de similarité qui est combinée cette fois ci, la formule de recommandation devient :

$$R_{a,k} = \frac{\sum_{t=1}^{K} \left(R_{a,t}.SimTot(i_k,i_t) \right)}{\sum_{t=1}^{K} SimTot(i_k,i_t)} \qquad (5)$$

Nous concluons que l'intégration sémantique apporte deux avantages pour les SFC, premièrement l'utilisation de ces information sémantiques améliorent les mesures de similarité et la comparaison entre objets ce qui augmente la précision de recommandation (intérêt majeur pour ces systèmes).
Deuxièmement, ces méta-données décrivent des nouveaux items intégrés au système ce qui réduit l'effet de démarrage a froid.

3. La technique LSI

Le modèle LSI (Latent Semantic Indexing) proposée par Derwester et al. [DER90], est un modèle algébrique de RI fondé sur la décomposition en valeurs singulières (*SVD : Singular Value Decomposition*) de la matrice (termes -documents) qui représente l'espace d'indexation du modèle vectoriel, cette matrice est projetée dans un espace de dimensions plus faible. Beaucoup d'applications RI ont montré que l'application de cette technique améliore la qualité de précision [MOB04].

Chaque dimension dans l'espace réduit à une variable latente (ou facteur) qui représente les indices des termes les plus corrélés, ainsi réduire la dimension de la matrice originelle permet de réduire la quantité de bruit dans les données et par conséquent diminue le temps de calcul ainsi que l'apaisement de la non évaluation de la matrice initiale, de ce fait, il y'a une amélioration sur le calcul de la similarité entre les documents indexés et les requêtes utilisateur.
Ici, nous appliquons cette idée pour créer un espace de dimensions réduit pour les attributs sémantiques associés aux items et sur la matrice d'évaluation (user-item) dans un autre cas.

Soit $S_{n\times d}$ la matrice sémantique de n items et d attributs, par décomposition SVD :

$$S_{n\times d}=U_{n\times r}\bullet \Sigma_{r\times r}\bullet V_{r\times d} \qquad (6)$$

où U et V sont deux matrices orthogonales (U.U'=I), r est le rang de la matrice S et Σ est une matrice diagonale de taille r × r, où sa diagonale contient toutes les valeurs singulières de la matrice S triées par ordre décroissant.
Il est prouvé qu'il existe une seule décomposition de cette manière [LOA96].
Un avantage de SVD est qu'il fournit la meilleure approximation de la matrice d'origine S avec rang inférieur [BER95].
Nous pouvons réduire le rang de la diagonale de la matrice Σ á un rang inférieur k (k<r) de façon á maintenir que les k plus grandes valeurs singulières, par conséquent on réduit U á U' et V á V' et la matrice approximative S devient :

$$S'_{n\times d}=U'_{n\times k}\bullet \Sigma'_{k\times k}\bullet V'_{k\times d} \qquad (7)$$

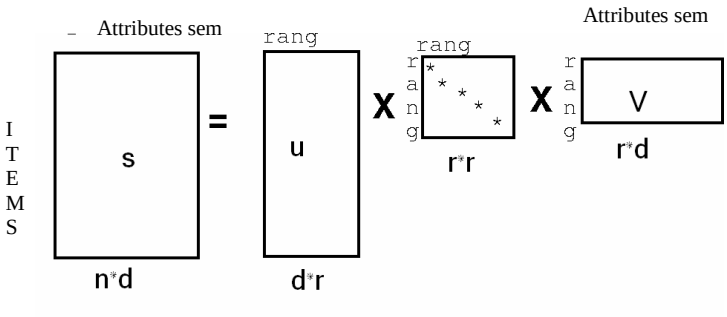


Fig III.3 - Décomposition en SVD de la matrice S

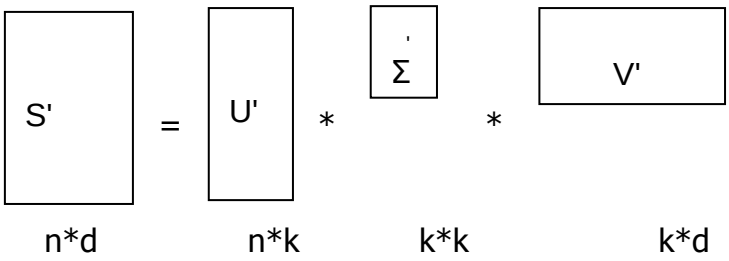


Fig IV.4 - Réduction de la SVD de la matrice S

Remarque : le paramètre k est important à définir, pas théoriquement mais par les tests empiriques. (K≈r : complexité de calcul, K<<r : perte d'information)

S' est la matrice sémantique réduite (moins creuse) améliore la similarité et réduire le processus de calcul.

En outre, les variables latentes générées représentent des groupes de forte corrélation des attributs dans les données d'origines, donc potentiellement réduire la quantité de bruit associée à l'information sémantique.

Comme conclusion, l'application de cette technique sur la matrice sémantique ou bien sur la matrice d'évaluation conduit á une progression de la qualité de précision et réduit le coût de calcul.

CHAPITRE IV

EXPÉRIMENTATIONS ET RÉSULTATS

Dans cette partie, nous présentons les résultats d'expérimentation effectués pour évaluer les différentes propositions du chapitre précédent. Les évaluations portent essentiellement sur la méthode du filtrage collaboratif proposée et les améliorations adoptées.
En premier lieu, nous décrivons le jeu de donnes MovieLens sur lequel on a effectué nos évaluations.
Par la suite, on se focalise sur les mesures utilisées pour analyser les résultats obtenus ainsi que l'approche suivie et nous finissons par une synthèse et une conclusion de ce chapitre.

1. Jeu de données

Pour expérimenter divers aspects du modèle proposé dans ce mémoire, un modèle basé items boosté par une infrastructure web sémantique, nous utilisons un jeu de données réel du système de recommandation de films MovieLens14 . Ces jeux de données sont très utilisés dans beaucoup d'études du domaine de filtrage collaboratif [HKR00].

Les données concernent 943 inscrits, 1682 titres de films et 100000 évaluations, les films sont notés sur une échelle de 1 à 5 et chaque utilisateur a noté au moins 20 films (entre 20 et 737) au total 93.7% des notes sont manquantes (matrice creuse).

Le système recommande des films de genres différents, le public visé est assez large quant à l'âge, le genre, la situation géographique et la profession, l'âge des utilisateurs oscille entre 7 et 73 ans.
La base contient 21 occupations différentes, parmi elles on trouve des métiers très variées : personne au foyer, étudiant, personne à la retraite, écrivain, scientifique, etc. Cela nous révèle un public très varié dont l'intérêt commun est le goût pour le cinéma.

Le jeu de données est constitué des tables suivantes :

● table rating

La table *rating* contient 100 000 évaluations données par les utilisateurs aux différents films (ITEMS) qui leurs ont été recommandé par le système, elle est formée de quatre colonnes :
-userid : identifiant de l'utilisateur
-movieid : identifiant du film
-score : le score donné par l'utilisateur au film, ce score peut prendre des valeurs entre $eval_{min}$= 0 et $eval_{max}$=5, un film est jugé pertinent par l'utilisateur s'il a un score supérieur ou égal à 4.
- date évaluation (temps cumul en seconde à partir du 01/01/1970).

14 http://movielens.umn.edu/

ID USER	ID MOVIE	RATING	DATE RATING
561	3	3	885810390
537	750	3	886028498
897	419	4	879990430
823	68	3	878438930
864	64	5	888887830
607	275	4	883879756
426	205	4	879444893
64	326	3	879365313
603	180	4	891956946
616	313	5	891224590
379	47	5	880740461
244	845	3	880606634
763	173	4	878914968
717	237	5	884642400
843	252	3	879445114
854	270	4	882811810
315	301	2	879799327
378	118	4	880044879
299	4	3	889503074

TAB IV.1- ÉVALUATION

- **Table movie**

La table movie contient 1682 enregistrements, elle est formée de cinq colonnes :
-id movie.
-title : le titre du film.
-release date: date de réalisation du film.
- url: lien vers des pages web concernant le movie.
-genre : le(s) genre(s) au(x)quel(s) le film appartient.

ID MOVIE	TITLE	RELEASE DATE	URL	GENRE
1	Toy Story (1995)	01-Jan-95	http://us.imdb.com/M/title-e	0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0
2	GoldenEye (1995)	01-Jan-95	http://us.imdb.com/M/title-e	0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0
3	Four Rooms (1995)	01-Jan-95	http://us.imdb.com/M/title-e	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0
4	Get Shorty (1995)	01-Jan-95	http://us.imdb.com/M/title-e	0 1 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0
5	Copycat (1995)	01-Jan-95	http://us.imdb.com/M/title-e	0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 1 0 0
6	Shanghai Triad (Yao a yao yao da	01-Jan-95	http://us.imdb.com/Title?Ya	0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0
7	Twelve Monkeys (1995)	01-Jan-95	http://us.imdb.com/M/title-e	0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0
8	Babe (1995)	01-Jan-95	http://us.imdb.com/M/title-e	0 0 0 0 1 1 0 0 1 0 0 0 0 0 0 0 0 0
9	Dead Man Walking (1995)	01-Jan-95	http://us.imdb.com/M/title-e	0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
10	Richard III (1995)	22-Jan-96	http://us.imdb.com/M/title-e	0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1
11	Seven (Se7en) (1995)	01-Jan-95	http://us.imdb.com/M/title-e	0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0
12	Usual Suspects, The (1995)	14-Aug-1995	http://us.imdb.com/M/title-e	0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0
13	Mighty Aphrodite (1995)	30-Oct-95	http://us.imdb.com/M/title-e	0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
14	Postino, Il (1994)	01-Jan-94	http://us.imdb.com/M/title-e	0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0
15	Mr. Holland's Opus (1995)	29-Jan-96	http://us.imdb.com/M/title-e	0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
16	French Twist (Gazon maudit) (199	01-Jan-95	http://us.imdb.com/M/title-e	0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0
17	From Dusk Till Dawn (1996)	05-Feb-1996	http://us.imdb.com/M/title-e	0 1 0 0 0 1 1 0 0 0 0 1 0 0 0 0 1 0
18	White Balloon, The (1995)	01-Jan-95	http://us.imdb.com/M/title-e	0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
19	Antonia's Line (1995)	01-Jan-95	http://us.imdb.com/M/title-e	0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
20	Angels and Insects (1995)	01-Jan-95	http://us.imdb.com/M/title-e	0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0
21	Muppet Treasure Island (1996)	16-Feb-1996	http://us.imdb.com/M/title-e	0 1 1 0 0 1 0 0 0 0 0 0 1 0 0 0 1 0
22	Braveheart (1995)	16-Feb-1996	http://us.imdb.com/M/title-e	0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0
23	Taxi Driver (1976)	16-Feb-1996	http://us.imdb.com/M/title-e	0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0
24	Rumble in the Bronx (1995)	23-Feb-1996	http://us.imdb.com/M/title-e	0 1 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
25	Birdcage, The (1996)	08-Mar-96	http://us.imdb.com/M/title-e	0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
26	Brothers McMullen, The (1995)	01-Jan-95	http://us.imdb.com/M/title-e	0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0

TAB IV.2 - MOVIE

● Table genre

La table genre contient 19 genres de films, elle contient deux colonnes :
-id genre: l'identifiant du genre
-genre : l'intitulé du genre.

ID GENRE	GENRE
0	unknown
1	Action
2	Adventure
3	Animation
4	Children's
5	Comedy
6	Crime
7	Documentary
8	Drama
9	Fantasy
10	Film-Noir
11	Horror
12	Musical
13	Mystery
14	Romance
15	Sci-Fi
16	Thriller
17	War
18	Western

TAB IV.3 - GENRE

Dans la table movie TABLE5.2 La valeur de l'attribut genre est mentionnée par 1 i.e la présence de ce genre dans le film ou 0 sinon.

● Table user

La table user contient 943 utilisateurs elles est formée de cinq colonnes :
-id user
-age
-genre : féminin ou masculin
-occupation
-Zipcode : le code de la ville de l'utilisateur.

ID USER	AGE	GENRE	OCCUPATION	ZIP CODE
1	24	M	technician	85711
2	53	F	other	94043
3	23	M	writer	32067
4	24	M	technician	43537
5	33	F	other	15213
6	42	M	executive	98101
7	57	M	administrator	91344
8	36	M	administrator	5201
9	29	M	student	1002
10	53	M	lawyer	90703
11	39	F	other	30329
12	28	F	other	6405
13	47	M	educator	29206
14	45	M	scientist	55106
15	49	F	educator	97301
16	21	M	entertainment	10309
17	30	M	programmer	6355
18	35	F	other	37212
19	40	M	librarian	2138
20	42	F	homemaker	95660
21	26	M	writer	30068
22	25	M	writer	40206
23	30	F	artist	48197
24	21	F	artist	94533
25	39	M	engineer	55107
26	49	M	engineer	21044
27	40	F	librarian	30030
28	32	M	writer	55369

TAB IV.4 -USER

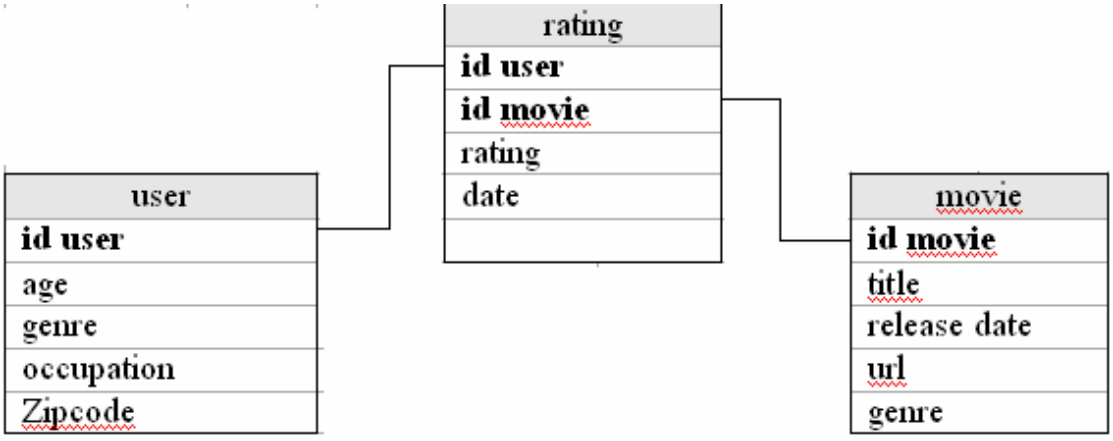


Fig IV.1 - Liens entre tables

2. Les mesures d'évaluation

Pour évaluer la précision, l'efficacité et la performance des algorithmes proposés (algorithme basé items, algorithme basé sur la sémantique des items et l'algorithme basé sur la technique SVD), nous employons la mesure MAE utilisée spécialement pour les systèmes de filtrage et ainsi que les mesures de précision, de rappel et de f-mesure (mesures utilisées dans le domaine de recherche d'information).

2.1 Le MAE

La mesure erreur moyenne absolue (MAE – Mean Absolute Error) [RAC02] calcule la différence moyenne entre les prédictions p_j calculées par le système et les scores e_j donnés réellement par l'utilisateur dans le processus de l'évaluation.

$$\overline{|E|} = \frac{\sum_{i=1}^{N} |e_i - p_i|}{N}$$

Où e_i : l'évaluation réelle du i^eme item par l'utilisateur, p_i : la prédiction du système du i^eme item pour cet utilisateur et N : le nombre des items évalués par l'utilisateur.

Bien sûr, plus le MAE sera petit, plus la performance du système est meilleure

2.2 Le rappel

Le rappel représente la proportion des items pertinents retournés par l'algorithme par rapport au nombre total des items pertinents existants. Sa formule est :

$$R = \frac{N_{pr}}{N_p}$$

N_{pr} : Le nombre des items pertinents renvoyés par l'algorithme,

N_p : Le nombre des items pertinents.

Le rappel mesure l'efficacité d'un algorithme. Plus la valeur de rappel est élevée plus le résultat de l'algorithme couvre toutes les items correctes.

2.3 La précision

La précision est la proportion des items pertinents parmi l'ensemble de celles renvoyés par l'algorithme.

$$P = \frac{N_{pr}}{N_r}$$

N_{pr} : Le nombre des items pertinents renvoyés par l'algorithme,

N_r : Le nombre des items renvoyés.

Plus la valeur de précision est élevée, plus le bruit dans le résultat de l'algorithme est réduit, et donc plus la qualité de résultat est meilleure.

Les systèmes tendent à améliorer le taux de précision et de rappel, le système « idéal » devrait avoir pour objectif un rappel et une précision élevés. Cependant ces deux taux ont souvent tendance à fonctionner en opposition, et chaque système cherche généralement un

équilibre entre ces deux valeurs, favorisant parfois l'une au dépend de l'autre selon le but visé.
La précision peut être directement calculée par les jugements de pertinence, le rappel est un peu plus compliqué parce qu'il ne dépend pas seulement de ce qui a été retourné mais aussi sur des documents qui sont pertinents et qui n'ont pas été retrouvés par le système.
Le plus souvent, on obtient un taux de précision et de rappel aux alentours de 30% [RIS'79].

2.4 La F-mesure

Est un compromis entre le rappel et la précision. Elle permet de comparer les performances des algorithmes par une seule mesure. La f-mesure est définie par:

$$F = \frac{2 * R * P}{R + P}$$

3. Démarche d'évaluation

3.1 Procédé global

Notre objectif dans ce mémoire est d'améliorer la qualité des systèmes de filtrage type collaboratif, on se bénéficiant de l'information sémantique extraite à partir des items pour renforcer les algorithmes classiques ensuite nous introduisons une technique LSI pour traiter l'inconvénient de la matrice creuse et le passage à l'échelle.

La démarche suivie pour l'évaluation dans cette section est analogue à celle présentée dans le chapitre proposition.
Nous avons tout d'abord implémenté l'algorithme du filtrage collaboratif basé items sur un jeu de test (environs 30% de la base totale) puis nous le comparons avec d'autres algorithmes classiques basés users.

En deuxième étape, nous avons exploité la table movie pour l'extraction de l'information sémantique, nous alimentons l'algorithme précédent par ces données et voir les résultats obtenus, une discussion sur la combinaison de ces deux procèdes s'avère utile.

Finalement et afin d'optimiser le système de recommandation nous avons appliqué une technique LSI (latent semantic indexing) détaillée dans la section § III.3 et voir les résultats de chaque paradigme basé items, sémantique et combiné.

3.2 Outils d'évaluation

Notre évaluation est basée essentiellement sur les fichiers de donnés réelles téléchargés du site MovieLens, cet ensemble de données est transformé sous formes de matrices telle que la matrice d'évaluation $E_{943\times1682}$ indiquant les scores d'évaluation des 943 users pour les 1682 items.
Nous avons utilisé l'outil MATLAB fortement orienté pour le traitement matriciel, la matrice $E_{943\times1682}$, la matrice $S_{1682\times1682}$ indiquant les valeurs de similarités entre items et enfin comme résultat de calcul on génère une matrice de prédiction $P_{943\times1682}$ indiquant les valeurs de prédiction selon l'approche adoptée.

Matlab (MATrixLABoratory produit par MathWorks[15]) est un langage scientifique interactif et un outil puissant, simple et efficace orienté au calcul matriciel (addition, multiplication,

[15] www.mathworks.com/.

inversion, décompositions, déterminants, etc.), avec une puissante librairie de visualisation graphique 2D et 3D.
Il existe deux modes de fonctionnement :
Mode interactif: MATLAB exécute les instructions au fur et à mesure qu'elles sont données par l'usager.
Mode exécutif: l'utilisateur peut lui-même définit ses propres fonctions, en regroupant des instructions MATLAB dans un fichier portant le suffixe ".m".
Notons enfin que MATLAB est disponible sur tous types de plates-formes (toutes les stations sous UNIX y compris LINUX, Windows 9x-xp et Macintosh).
(MATLAB ANNEXE A).

4. Résultats

4.1 Algorithme basé évaluation

a) La mesure MAE

. La matrice initiale : $E_{200\times400}$ (200 utilisateurs et 400 items)

. La taille de voisinage : 10 à 90.

```
55    5   0   0   0   0   0   3   0   0   0   0   0   5   0   4   0   0   0   0
56    0   0   0   0   0   0   4   0   0   0   0   0   0   0   0   0   0   0   0
57    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
58    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
59    2   1   3   2   0   0   4   3   0   3   3   4   3   0   0   0   2   0   0
60    0   0   0   0   0   0   0   0   4   0   0   0   0   0   2   0   0   0   0
61 -  [lig,col]=size(E);
62 -  S=eye(col);
63 -  for j=1:col-1
64 -      v1=E(:,j);
65 -      for k=j+1:col
66 -          v2=E(:,k);
67 -          inc1=find(v1 & v2);
68 -          x=(sum(v1(inc1).*v2(inc1)))/(sqrt(sum(v1(inc1).^2)*sum(v2(inc1).^
69 -          S(j,k)=x;
70 -          S(k,j)=x;
71 -      end ;
72 -  end;
73    %disp('la matrice devaluation est: ');
74 -  E ;
75    %disp('la matrice de similarite resultante est : ');
76 -  S;
77 -  P=E-E;
78 -  ii=1;
79 -  vp2=randnc(1,col-10);
80 -  for nbv=10:10:col-10
81 -      vp1=[10:10:col-10];
82 -  for i=1:lig
83 -      for j=1:col
84 -          var1=0;
85 -          var2=0;
```

Fig IV.2 - Algorithme basé évaluation

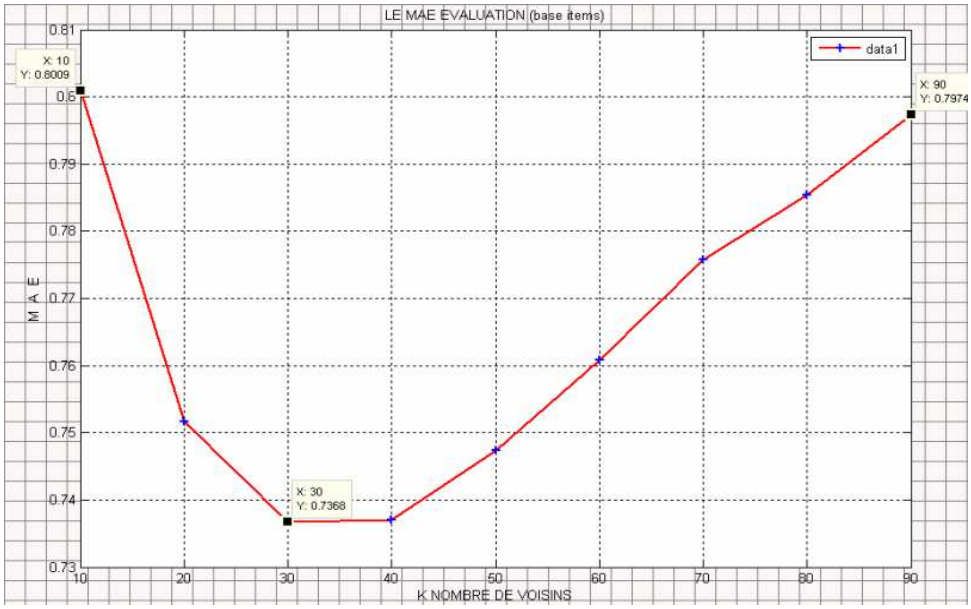


Fig IV.3 -MAE par l'algorithme basé évaluation

La figure IV.3 montre clairement que le MAE passe de 0.8009 pour 10 voisins à la valeur optimale 0.7368 au voisinage de 30 à 40 voisins, ensuite l'erreur augmente proportionnellement à l'augmentation du nombre de voisins ce qui traduit logiquement que la similarité se dégrade entre items à partir d'un rang donné (>40 voisins) et par conséquent l'augmentation automatique de l'erreur (MAE).

b) Les mesures rappel, précision et F-mesure

. La matrice initiale : $E_{200\times400}$ (200 utilisateurs et 400 items)

. La taille de voisinage : 10 à 200.

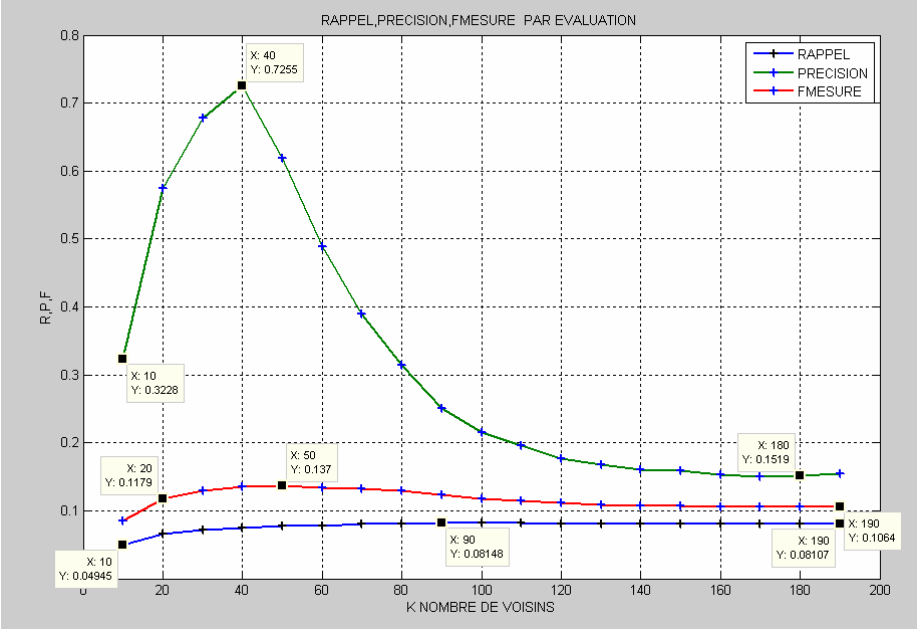


Fig IV.4 - Rappel, précision et f_mesure par l'algorithme évaluation

Au voisinage de 20 á 60 voisins nous constatons que le rappel atteint 8.1% Cette valeur signifie qu'un faible d'items pertinents est retourné par l'algorithme et cela dû aux taux d'information manquant dans la matrice initiale, ainsi qu'une précision allant jusqu'á72.5% ce qui signifie une bonne qualité de prédiction et un taux de bruit très faible, ces optimums sont obtenus a cause de la forte similarité entre items dans l'intervalle 20-60 voisins. La F-mesure comme une fonction linéaire entre le rappel et la précision est assez bonne, oscille dans l'intervalle [0.08 – 0.137].

c) Méthodes classiques basées users

Afin de voir l'impact de notre approche basée item énoncée au chapitre précédent et la comparer avec d'autres approches plus anciennes, nous avons implémenté deux autres algorithmes classiques du filtrage collaboratif basés users:

● Vote "Maximum"

Attribuer la note majoritaire (maximum) des k voisins (users) sur l'item j pour l'utilisateur actif i

$$P(i, j)=MAX_k P(k, j)$$

● Vote "Moyen"

Attribuer la note moyenne des k voisins (users) sur l'item j pour l'utilisateur actif i

$$P(i, j)=\frac{\sum_k P(k, j)}{k}$$

. La matrice initiale : $E_{200\times400}$ (200 utilisateurs et 400 items)
. La mesure : MAE.
. La taille de voisinage : 10 à 100.

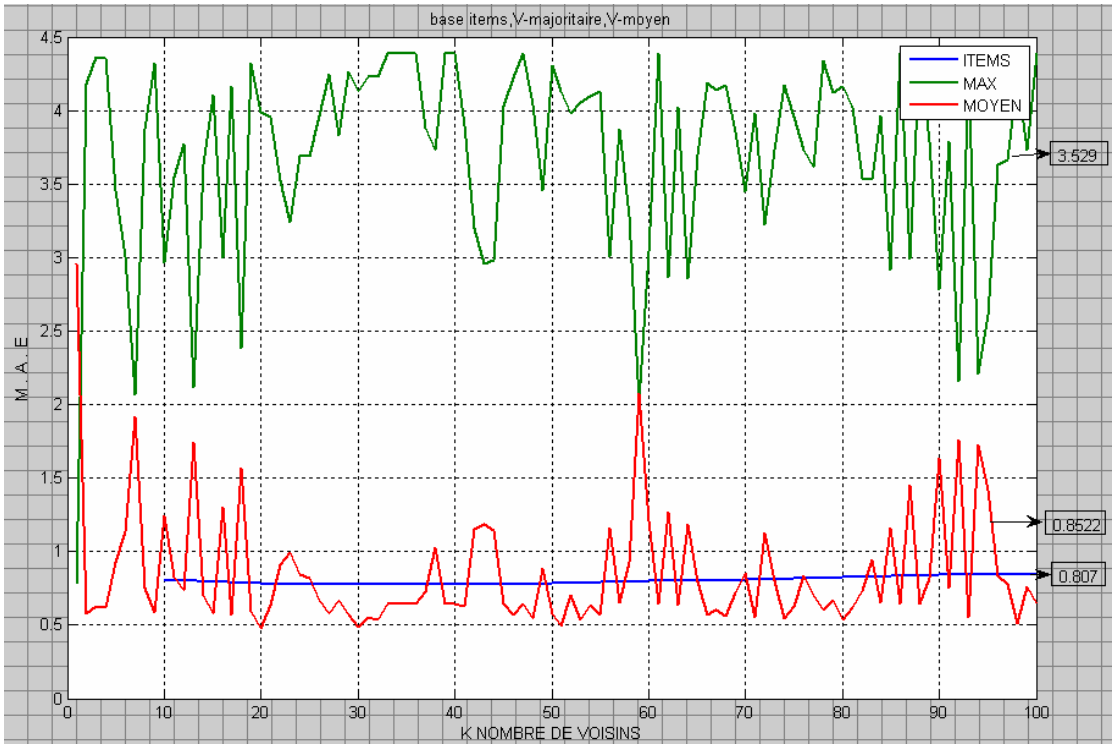


Fig IV.5 - M A E (vote moyen, vote maximum, items)

Selon les courbes obtenues dans la figureIV.5, on constate que la méthode par vote maximum ne donne pas un résultat important, une erreur moyenne de 3.529. Par contre la méthode par vote moyen a donnée un MAE de 0.8522 et notre méthode proposée basée items donne un MAE de 0.807, un résultat relativement meilleur par rapport au deux autres méthodes.

4.2 Algorithme sémantique

. La matrice initiale : $ES_{500\times20}$. (500 items en lignes et 20 attributs en colonnes)
. La mesure : MAE, rappel, précision.
. La taille de voisinage : 10 à 100.

L'algorithme est basé uniquement sur la matrice sémantique ES inspirée de la table item du ML-dataset (TABLEIV.2).

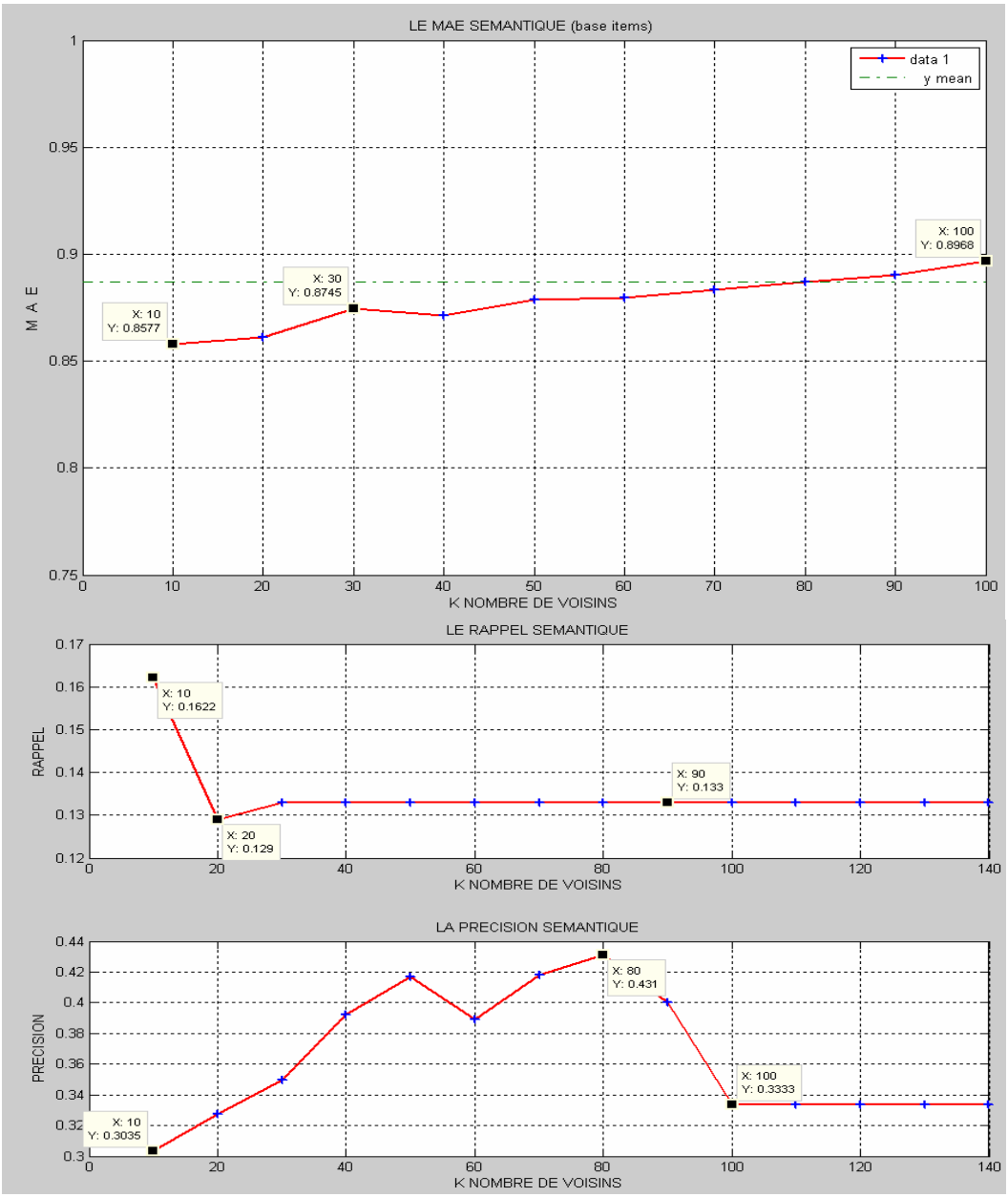


Fig IV.6 - Algorithme sémantique

Malgré le nombre restreint d'attributs, on constate un résultat satisfaisant.
D'après la figureIV6 un MAE variant de 0.8577 au 0.8968 et un rappel atteint 16.22% et un taux de précision de 43.1%.

	MAE	Rappel	Précision
Algorithme par évaluation	0.7368 - 0.8009	8.1%	72.5%
Algorithme sémantique	0.8577 - 08968	16.22%	43.1%

Tab IV.5 - Comparaison des algorithmes par évaluation et sémantique

4.3 Algorithme hybride

. La matrice initiale : $ES_{500\times20}$, $E_{200\times400}$.

. La mesure : MAE, rappel, précision.

. La taille de voisinage : 10 à 90.

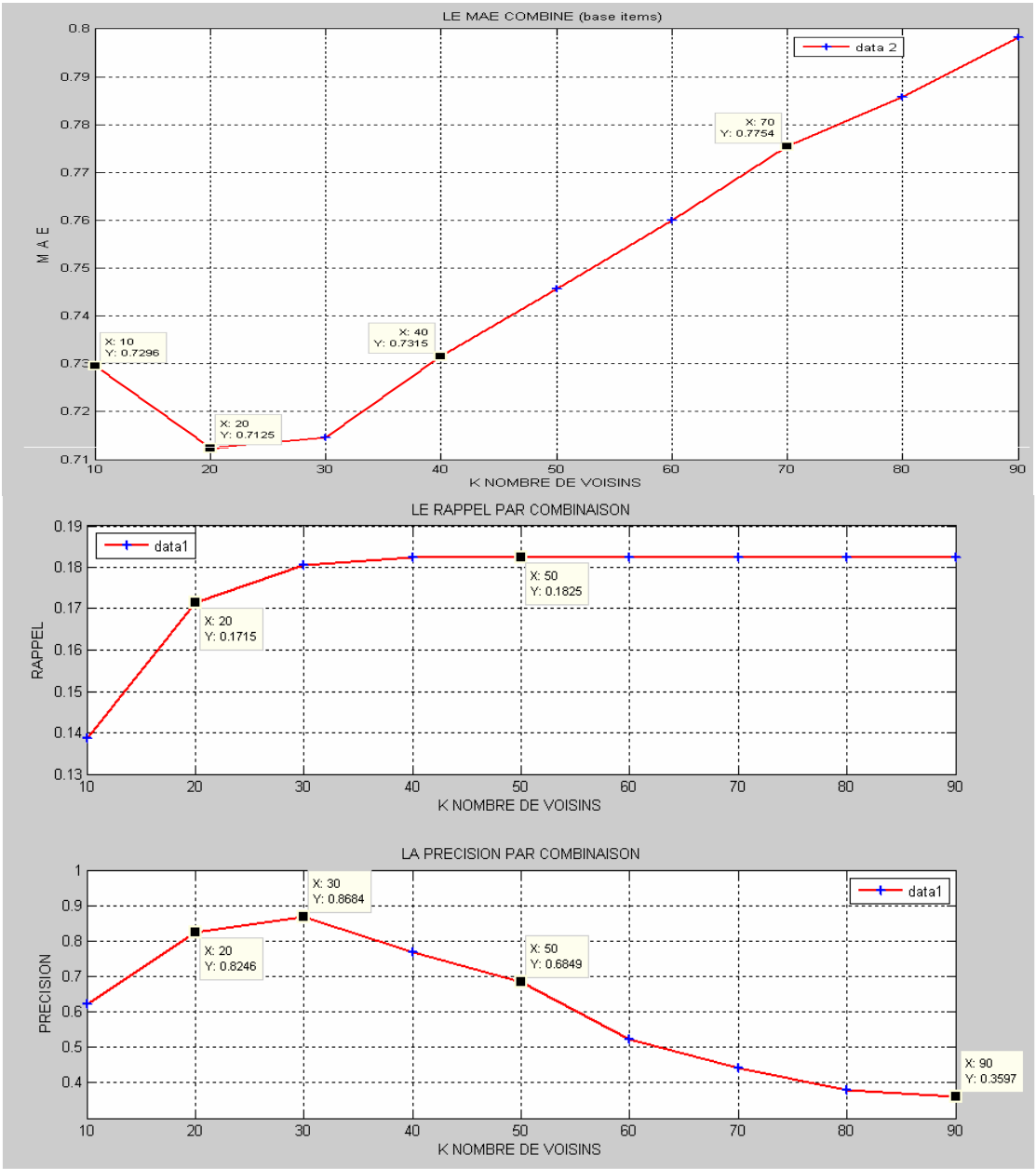


Fig IV.7 - Algorithme hybride

La combinaison des deux algorithmes nous a donné des bon résultats, un MAE=0.7125 pour les 20 voisins les plus similaires et un taux de rappel atteint 18.25% et une précision allant jusqu'à 86.84% (figure-IV.7).

Donc l'intégration de l'information sémantique a permet de diminuer l'erreur de prédiction et augmenter le taux de rappel et de précision.

4.4 Optimisation par LSI

Sachant que le taux des évaluations communes entre items est trop faibles ce qui implique une matrice creuse (voir les matrices E, ES), de cet effet la similarité entre les objets (items dans notre cas, users dans les CF classiques) est faible ce qui influe directement sur la qualité de prédiction, pour remédier à ce problème on introduit une technique LSI (latent semantique indexing).

4.4.1 Algorithme basé évaluation

. La matrice initiale : $E_{200\times400}$ (200 utilisateurs et 400 items)

. La mesure : MAE, rappel, précision.

. La taille de voisinage : 10 à 200.

. SVDk : k=10

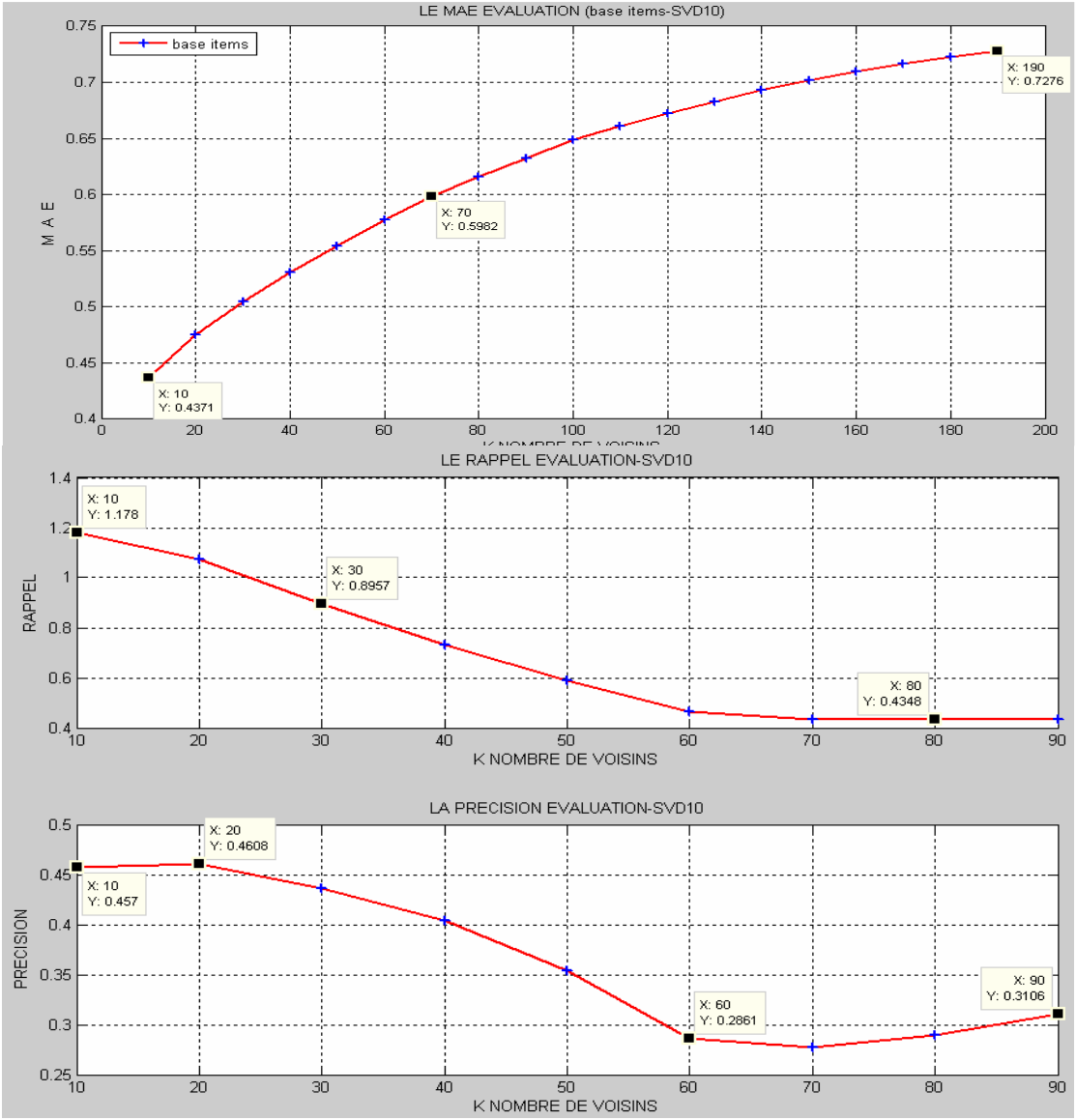


Fig IV.8 - Algorithme évaluation SVD10

Par l'application de la technique LSI on a constaté une amélioration remarquable (MAE entre 0.4371 et 0.7276, un rappel entre 117.8% et 43.48% et une précision entre 46.08% et 28.61%) par rapport aux résultats obtenus par l'algorithme basé évaluation (figureIV.3 et figureIV.4).

L'application du LSI allége l'effet de la matrice creuse ce qui produit une forte corrélation entre les items.

4.4.2 Algorithme sémantique

. La matrice initiale : $ES_{500\times20}$. (500 items et 20 attributs)

. La mesure : MAE, rappel, précision.

. La taille de voisinage : 10 à 190.

. SVDk : k=10

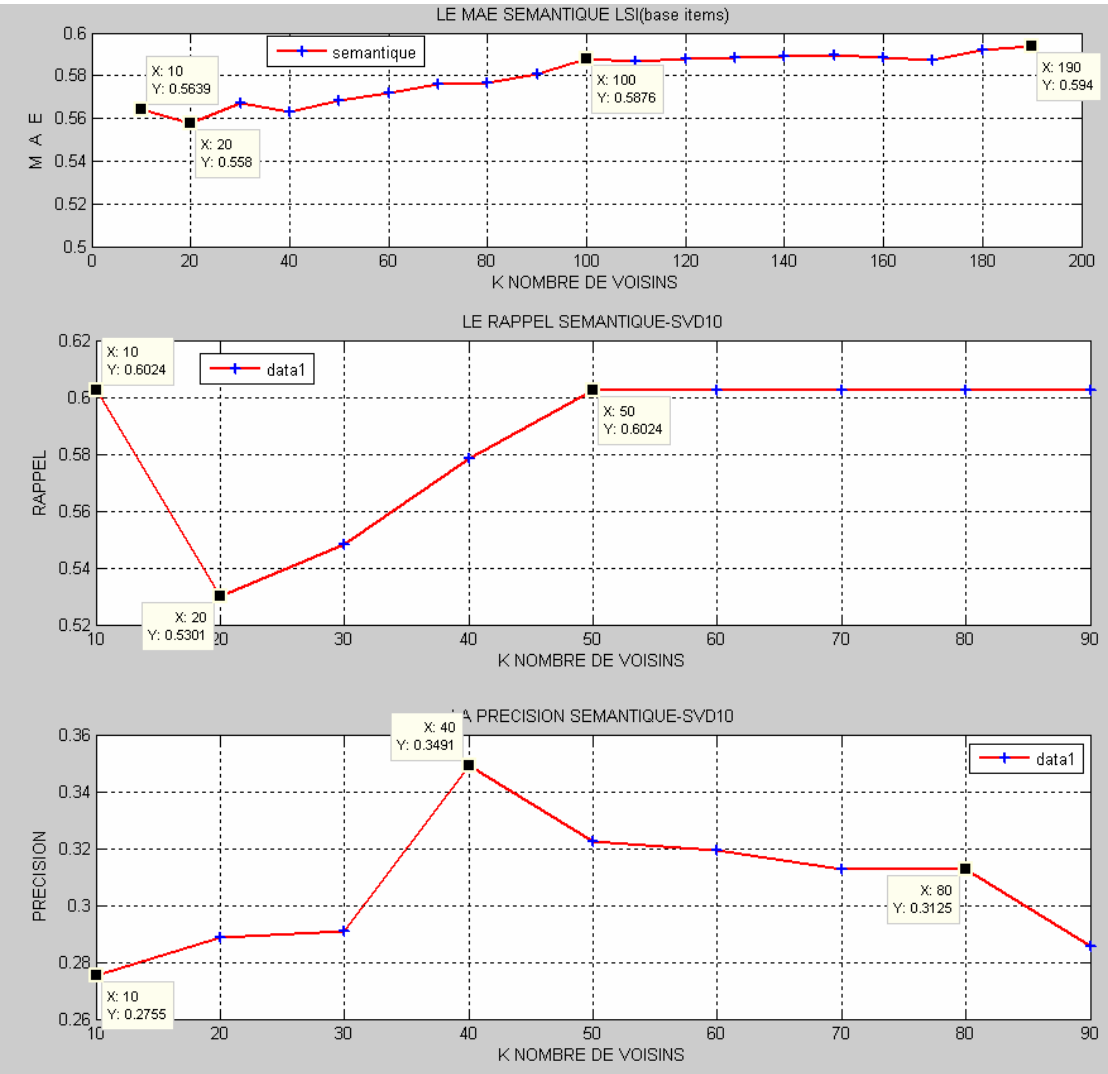


Fig IV.9 - Algorithme sémantique SVD10

Comparaison entre l'algorithme sémantique et sémantique par LSI

	MAE	Rappel	Précision
Algorithme sémantique	0.8577-0.8968	12.9%-16.22%	33.33%- 43.10%
Algorithme sémantique par LSI	0.5639-0.558	53.01%-60.24%	27.55%-34.91%

Tab IV.6 - Algorithme sémantique et LSI sémantique

Par les mêmes raisons citées ci-dessus, on constate que l'application de la technique LSI apporte une amélioration importante sur toutes les mesures, l'erreur moyenne absolue, le rappel et la précision.

4.4.3 Algorithme hybride

. La matrice initiale : $ES_{500\times20}$, $E200_{\times400}$.

. La mesure : MAE, rappel, précision.

. La taille de voisinage : 10 à 190.

. SVDk : k=10

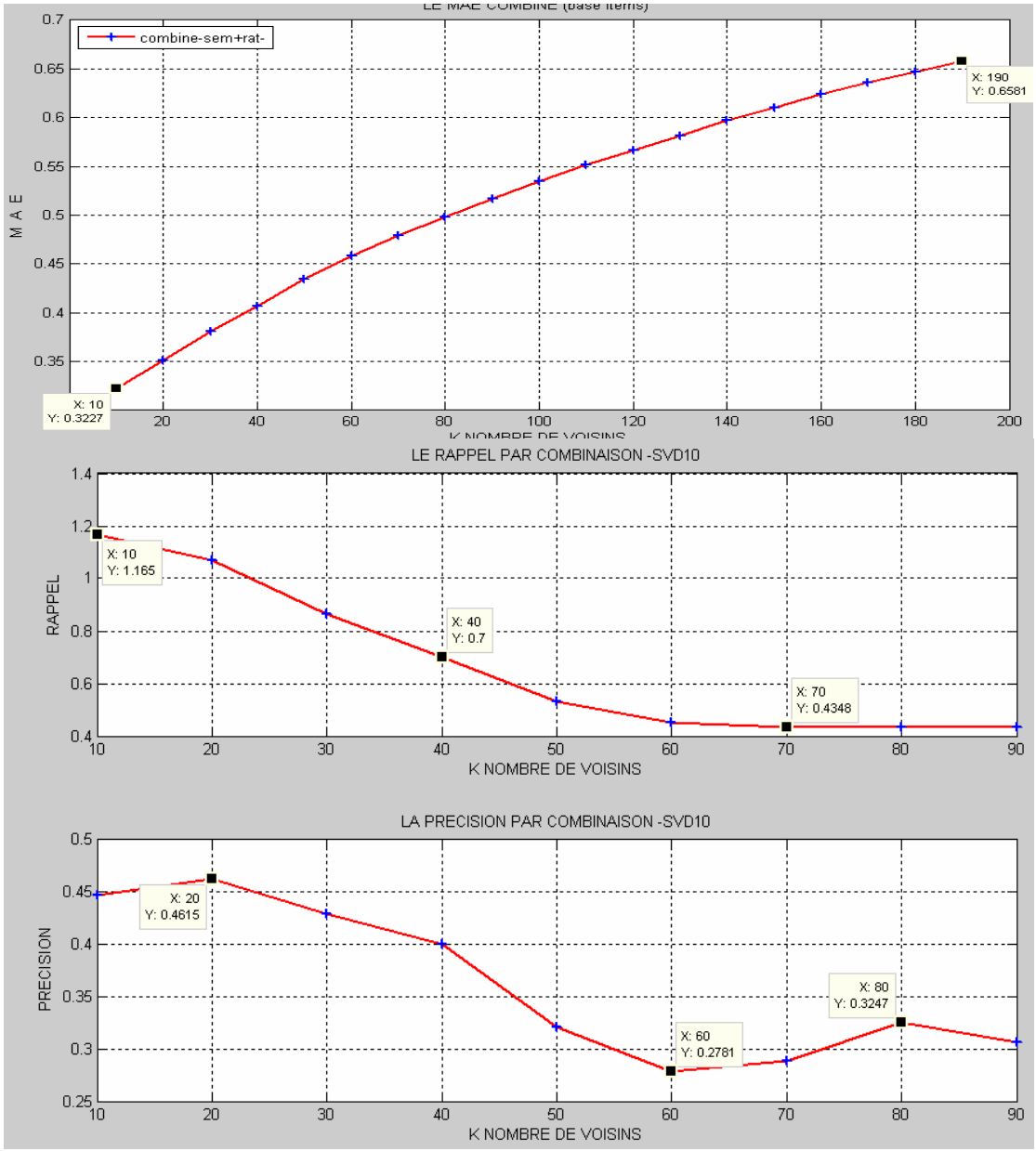


Fig IV.10 - Algorithme hybride-SVD10

La combinaison de l'algorithme par évaluation et l'algorithme sémantique nous a donné des bons résultats, l'introduction de la technique LSI a encore améliorée ces résultats

(MAE=0.3227-0.6581, rappel 43.48%-116.5%, précision =27.81%-46.15%) (figIV.9 et fig IV.10)

4.5 Résultats récapitulatifs

ALGORITHMES	Les mesures		
	M A E	**RAPPEL**	**PRECISION**
Evaluation	0.7368-0.8009	4.94%-8.14%	15.19%-72.55%
Sémantique	0.8577-0.8968	12.9%-16.22%	33.33%- 43.10%
Hybride (evel+sém)	0.7125-0.7985	13.98%-18.25%	35.97%-86.84%
évaluationSVD	0.4371-0.7276	117.8%-43.48%	28.62%-46.08%
sémantiqueSVD	0.5639-0.5580	53.01%-60.24%	27.55%-34.91%
hybrideSVD(evalsvd+semsvd)	0.3227-0.6581	116.5%-43.48%	46.15%-27.81%

Tab IV.7 - Résultats des différents algorithmes

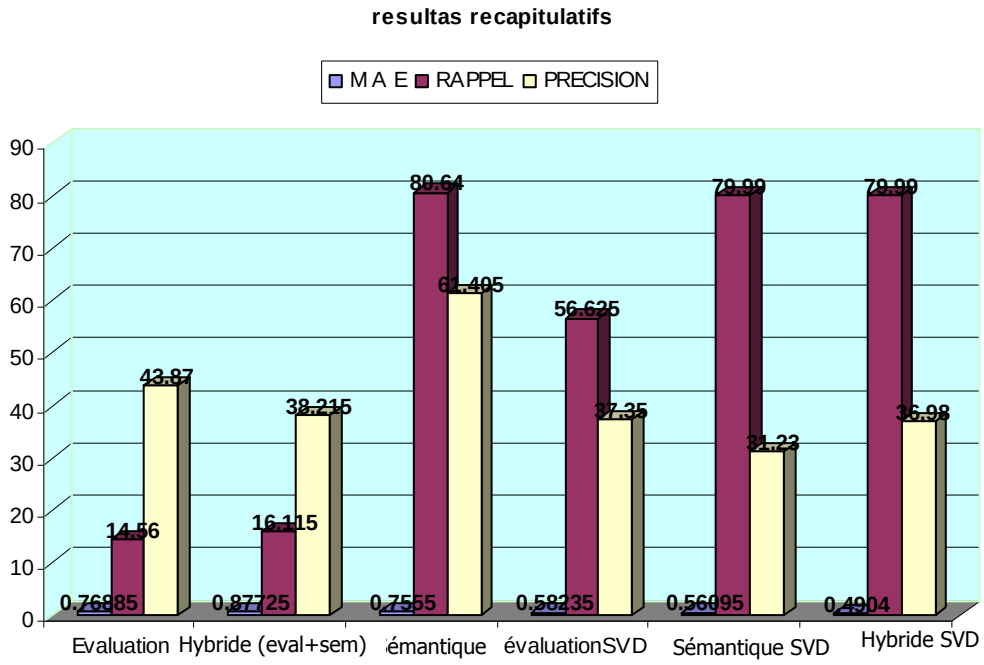


Fig IV.11 - Résultats des différents algorithmes

En conclusion, l'ajout de l'information sémantique et l'introduction des techniques d'optimisation comme la technique LSI et ainsi une hybridation entre ces approches ont menés à des résultats très satisfaisants (MAE=0.3227) et par conséquent une amélioration de la qualité des systèmes de filtrage collaboratif (SFC).

4.6 D'autres critères d'évaluation

- Le temps d'exécution en fonction de la taille de la matrice

. La matrice initiale : $E_{200\times300}$.

. La mesure : le temps d'exécution en secondes.

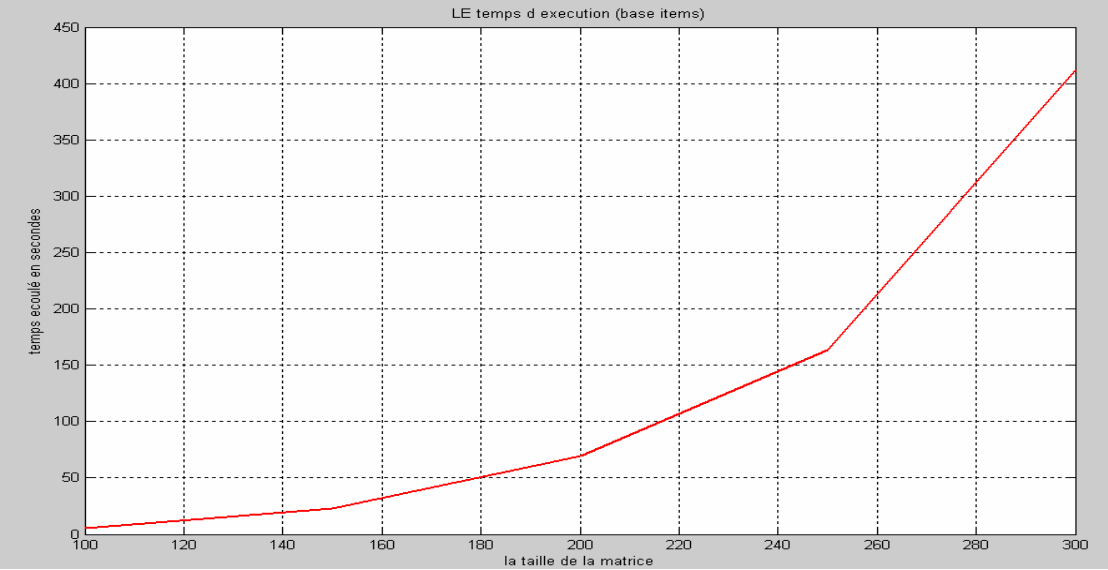


Fig IV.12 - Temps d'exécution

Bien évident, en examinant la figureIV.12 nous constatons que le temps d'exécution augmente proportionnellement par rapport à la taille de la matrice, un autre atout de l'approche basée items c'est le mode off-line i.e que le calcul de similarité entre items est indépendant du fonctionnement du système, cette tâche pourra être différée en temps opportun.

- Les variations du SVD

. La matrice initiale : $E_{400\times400}$.

. La mesure : pourcentage d'information préservé

. Variation du paramètre k du SVD

```
919    499 1958    0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0;
920    500 1996    0   0   1   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0];
921    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
922    %%%%  la decomposition de la matrice semantique ES SVD-10     %%%%%%%%
923    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
924
925 -  ES=ES(1:200,:);
926 -  ES=ES';
927 -  [u s v]=svds(ES,10);
928 -  ES=u*s*v';
929 -  [lig1,col1]=size(ES);
930
931    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
932    %%%%  la decomposition de la matrice evaluation E SVD-10     %%%%%%%%%
933    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
934 -  [u s v]=svds(E,10);
935 -  E=u*s*v';
936 -  [lig,col]=size(E);
937 -  SS=eye(col1);      % similarite semantique;
938 -  for j=1:col1-1
939 -      v1=ES(:,j);
```

Fig IV.13 - listing de la décompositionSVD10

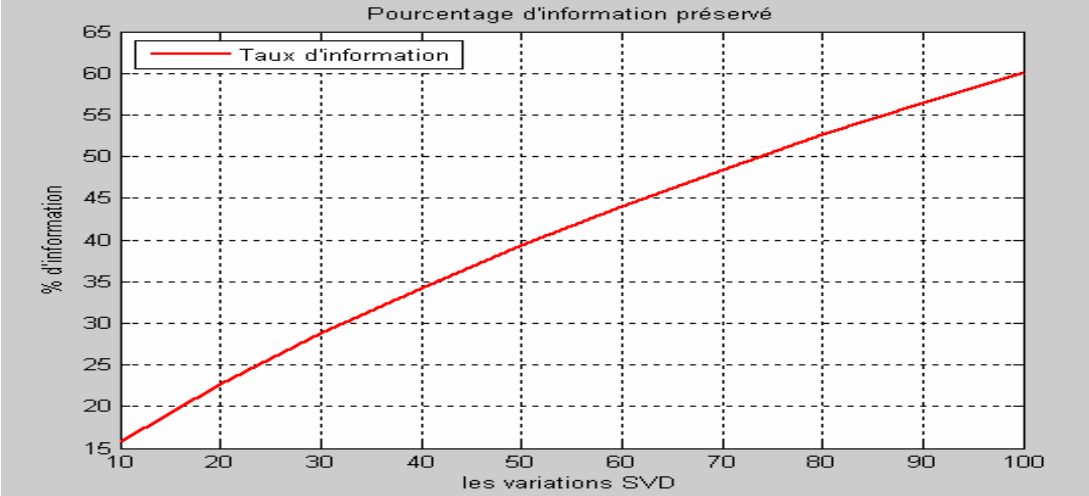


Fig - IV.14 variation du paramètre k du SVD

b)

. La matrice initiale : $E_{400\times400}$.

. SVDk: k=10,15,20,30,40,60.

. La mesure : MAE

. Variation du K voisins

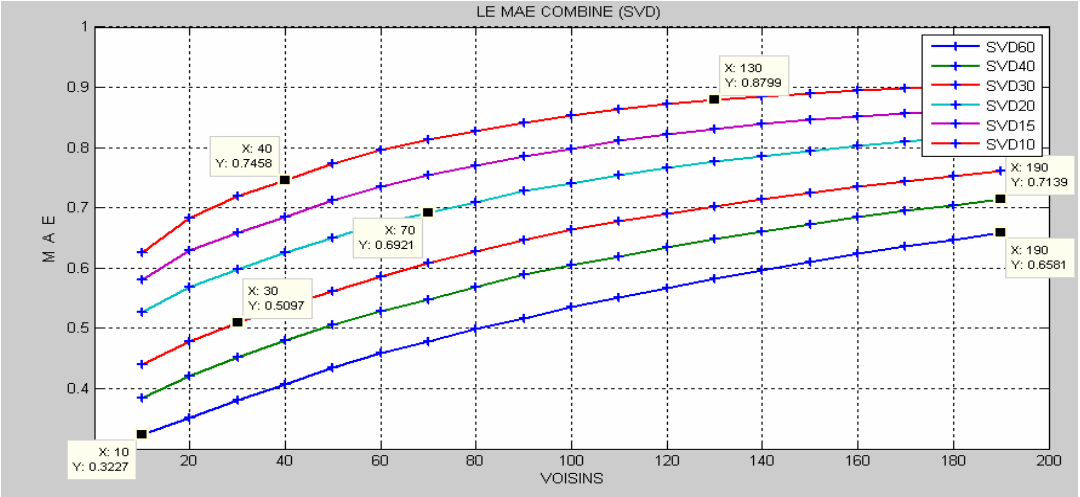


Fig IV.15 - MAE par SVD-k

L'introduction de la technique LSI a montré une performance impressionnante selon les tests empiriques effectués, d'un coté à réduire la complexité combinatoire i.e remédier au problème du passage à l'échelle, point discuté dans le chapitre précédent et d'un autre coté, le travail avec une matrice approximative moins creuse que la matrice originale améliore la qualité de prédiction.

5. Synthèse

Nous avons présenté dans cette dernière partie les résultats d'expérimentations de nos contributions qui sont basées essentiellement sur:

- La standardisation de l'approche basée items au lieu de rester figé uniquement sur l'utilisateur.
- D'un autre coté l'intégration de l'information sémantique nous a permis d'augmenter la mesure de similarité. On se base sur cette masse d'information en cas d'absence de l'information habituelle explicite ou bien on combine ces deux types d'information afin d'atteindre une meilleure qualité de prédiction. Initialement le système de filtrage peut exploiter cette information pour la recommandation et réduire l'effet de démarrage à froid.
- Finalement, et afin d'optimiser l'algorithme de prédiction on a appliqué la technique LSI pour faire le traitement sur des matrices moins creuses que celles utilisées à l'origine (trop d'information manquante), cette solution donne des résultats satisfaisants.

Par critères d'efficacité et de simplicité de l'outil MATLAB que nous avons utilisé, les résultats obtenus sont encourageants par rapport aux résultats antérieurs dans le domaine de filtrage collaboratif, ainsi ces résultats peuvent être améliorés par l'extraction efficace et optimale de l'information explicite et sémantique.

Conclusion et perspectives

Les systèmes de filtrage d'information fournissent à leurs utilisateurs un flot continu de documents, sans qu'ils aient à exprimer explicitement ce qu'ils cherchent. En contrepartie, il faut que ces systèmes connaissent le « profil »des utilisateurs, et suivent leurs évolution au cours du temps.
Un système de filtrage collaboratif (SFC) permet la découverte des items intéressants, grâce à l'automatisation du processus de la recommandation. Parmi les avantages offerts aux utilisateurs, on trouve :
- la possibilité d'exprimer leur avis quant à la pertinence des items, selon leurs goûts et la qualité qu'ils perçoivent sur ces items ;
- la possibilité de recevoir des recommandations inattendues car il suffit qu'un utilisateur, de profil proche, les ait jugées intéressantes ;
- la possibilité de bénéficier des évaluations sur les items que d'autres utilisateurs membres de sa communauté ont déjà faites.

Tous ces avantages sont apportés aux utilisateurs par le principe de collaboration, en contrepartie d'un effort individuel. Ces systèmes requièrent la participation active des utilisateurs, sur le long terme, pour atteindre de bonnes performances.
Les systèmes de recommandation ne se limitent pas à gérer des références à des items ou documents, mais supportent des domaines larges : cinéma, cuisine, assistance juridique, technologies de pointes, etc.

Dans ce mémoire, Nous avons mis l'accent sur l'approche basée items caractérisée par le calcul en mode off-line ainsi que la modularité des algorithmes de ce type d'approche, puis nous l'avons le amélioré en intégrant des informations sémantiques structurées sur les items pour des calculs de similarité.

Finalement, l'application de la technique latent semantic indexing basée sur la réduction de la décomposition en valeurs singulières de la matrice originelle, de ce fait allégeant le bruit des recommandations et par conséquent augmentant la précision du système et réduit la complexité combinatoire .

Nos résultats expérimentaux prouvent que l'approche basée items munie de l'infrastructure sémantique augmente et améliore l'information recommandée tout en traitant l'effet de démarrage à froid et le problème du manque de l'évaluation.

Nous prévoyons l'exploitation efficace et optimale des données sémantiques tenant en compte la structure de l'ontologie du domaine pour améliorer la pertinence de l'information recommandée.

Il semble utile d'étudier le paramètre de combinaison entre l'information sémantique (implicite) et l'information expliquée par les utilisateurs (explicite) selon les exigences et l'information disponible.

L'intégration des agents intelligents pour le traitement de la base profil et le traitement de l'information extérieure (corpus) augmente l'automatisation et la rapidité du processus de filtrage.

Bibliographie

[AGG99] Aggarwal, Charu C. Joel L. Wolf, Kun-Lung Wu, Philip S. Yu, “Horting Hatches an Egg: A New Graph-Theoretic Approach to Collaborative Filtering”, 1999.

[AMO08] Amokrane BELLOUI, NOUALI O. L’usage des concepts du web sémantique dans le filtrage d’information collaborative, thèse de magistère INI-Alger 2008.

[AMS99] Amato G., Straccia U., User Profile Modeling and Applications to Digital Libraries, In: Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries, Paris, France, 1999

[AUS04]. N. Aussenac-Gilles, J. Mothe, Ontologies as Background Knowledge to Explore Document Collections, In Actes de la Conférence sur la Recherche d'Information Assistée par Ordinateur (RIAO), pp 129-142, 2004.

[BAC06] BACH Thành Lê, Construction d’un Web sémantique multi-points de vue These de doctorat L’École des Mines de Paris à Sophia Antipolis octobre 2006

[BAE99] R. BAEZA-YATES , B. RIBEIRO-NETO , « Modern Information Retrieval » , ACM press et
Addison Wesley , New york , janvier 1999, http : //www.acm.org consulté le 03.08.2005

[BBZ95]. J. Bouaud, B. Bachimont, J. Charlet, and P. Zweigenbaum. Methodological principles for structuring an ontology. In ACM Press, editor, Proceedings of IJCAI’95 Workshop: Basic Ontological Issues in Knowledge sharing, 1995.

[BEC01] Bechhofer, S.; Horrocks, I.; Goble, C.; and Stevens, R., “OILEd: a Reason-able Ontology Editor for the Semantic Web”, In *Proceedings of KI2001, Joint German/Austrian conference on Artificial Intelligence*, September 19-21, Vienna. Springer-Verlag LNAI Vol. 2174, pp 396--408. 2001.

[BEL92] Belkin N.-J., Croft W.-B., Information Filtering and Information Retrieval: Two Sides of the Same Coin, *Communications of the ACM*, vol. 35 (12), 1992, p. 29-38.

[BER03] C. Berrut, “Filtrage collaboratif”, chapitre 8, p 255-283, E. GAUSSIER, M.H. STEFANINI, “Assistance intelligente à la recherche d’information”, Hermes-Lavoisier, 2003.

[BHK98] Breese J.-S., Heckerman D., Kadie C., Empirical Analysis of Predictive Algorithms for Collaborative Filtering, *Proceedings of the 14th Conference on Uncertainty In Artificial Intelligence (UAI’98)*, Wisconsin, USA, 1998

[BOK06] BOUZEGHOUBE, KOSTADINOF, Personnalisation de l'information: aperçu de l'état de l'art etdéfinition d'un modèle flexible de profils -2006

[BOU05] M. Boughanem, W. N. Zemirli, L. Tamine, and. Accès personnalisé à l'information : vers la dé_nition d'un pro_l utilisateur multidimensionnel . In International Symposium On Programming
Systems (ISPS) , Alger, 09/05/05-11/05/05, pages 20_28. USTHB, mai 2005.

[BOU07] M. Boughanem S. Walker, S. Robertson, , G. Jones, and K. S. Jones. Okapi at TREC-6 automatic and ad hoc, VLC, routing, filtering and QSDR. In Proceedings of TREC-6, pages 125–136.

[BOZ04] M. Bouzeghoub, “Action spécifique sur la personnalisation de l’information”, Laboratoire PriSM, Université de Versailles, CNRS-AS98/RTP9, France, 2004

[BRE98] Breese, J. S.; Heckerman, D.; Kadie, K.: Empirical analysis of predictive algorithms for collaborative filtering. Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI), 1998

[BUK02] Burke R. Hybrid Recommender Systems: Survey and Experiments, Journal of Personalization Research, User Modeling and User-Adapted Interaction, vol. 12 (4), p. 331-370, Kluwer Academic Publishers, 2002.

[CHA03] Charlet J. : L’ingénierie des connaissances. Développements, résultats et perspectives pour la gestion des connaissances médicales. Habilitation à diriger

Bibliographie

des recherches, Université Pierre et Marie Curie, 2003.

[COC06] An Te NGUYEN , COCoFil2 : Un nouveau système de filtrage collaboratif basé sur le modèle des espaces de communautés DOCTEUR DE L'UNIVERSITÉ JOSEPH FOURIER – GRENOBLE I 2006

[DCM03] Dublin Core Metadata Initiative. *Dublin Core Metadata Element Set, Version 1.1*. [en-ligne] Recommendation. Juillet, 2003. Disponible sur : <http://dublincore.org/documents/2003/07/02/dces/>

[DER90] Deerwester, S.T. Dumais, G.W. Furnas, T. K. Landauer, et R. Hrashman. Indexing by latent semantic analysis. *Journal of th american society for information science*, 41(6) :391–407, 1990.

[DES02]. E. DESMONTILS, C. JACQUIN, L. SIMON Vers un système d'annotation distribute 2002

[DOM98] Domingue, J., "Tadzebao and Webonto: Discussing, Browsing and Editing Ontologies on the Web", In *Proceedings of the Eleventh Knowledge Acquisition Workshop, KAW98*, Banff, 1998.

[ELH97] M. EL HACHANI, « L'indexation automatique », Note de Synthèse dirigée par M. HASSOUN, Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques (ENSSIB), Mars 1997

[EUZ02]. J. Euzenat, Eight questions about semantic Web annotations, IEEE Intelligent systems 17(2), pp 55-62, 2002.

[EUZ04]. Euzenat, J. et Valtchev, P. Similarity-based ontology alignment in OWL-lite. Dans Proc. 15th ECAI, Valencia (ES), 2004.

[FAR96] Farquhar; Fikes, R.; and Rice, J., "The Ontolingua Server: A Tool for Collaborative Ontology Construction", In *Proceedings of the 10th Knowledge Acquisition for Knowledge-Based Systems Workshop*, Banff, Alberta, Canada, 44.1-44.19, 1996, http://www.ksl.stanford.edu/software/ontolingua/].

[FLU95] Fluhr C., 1995, *Indexation et recherche d'information textuelle*, Ingénierie des langues, Hermes

[FRE04] Fresnel A., Pouliquen B., Riou C., Delamarre D., Le Beux P., 2004, *Computer Assisted Medical Diagnosis* - European Congress of the Internet in Medicine, Hospital.European Congress of the Internet in Medicine, Brighton 17

[GAL05] Gallardo-López M.-L., Accès à l'Information par un Système de Filtrage Collaboratif Contrôlé, Thèse, Université Joseph Fourier, Grenoble, France, 2005.

[GOL98]. Goldberg D., Oki B., Nichols D., Terry D.-B., Using Collaborative Filtering to Weave an Information Tapestry, *Communications of the ACM*, vol. 35 (12), 1992, p. 61-70.

[GOW03] J.P Mc Gowan. A multiple model approach to personnalised information access. In *Master thesis in computer science*. Faculty of science, University college Dublin, February 2003.

[GOW03] Gowan. J.P A multiple model approach to personnalised information access. In *Master thesis in computer science*. Faculty of science, University college Dublin, February 2003.

[GRU93] Gruber, T. R. (1993). Towards principles for the design of ontologies used for knowledge sharing. International Journal Human-Computer Studies, 43 :907_928.

[GUH03]. R.V. Guha, R. McCool, E. Miller, Semantic search, In Proceedings of the 12th International World Wide Web Conference, pp 700-709, 2003.

[GUI05]. Guarino. *Some Organizing Principles for a Unified Top-Level Ontology*. Revised version of a paper appeared at AAAI Spring Symposium on Ontological Engineering, LADSEBCNR
Int. Rep., February 2005.
http://www.ladseb.pd.cnr.it/infor/Ontology/Papers/OntologyPapers.html

[HAI05]. Ollivier Haemmerlé. *Systèmes Multi-Agents et Graphes Conceptuels : la plate-forme CoGITo.* Actes Actes des 3èmes Journées Francophones IAO et SMA, Saint Baldoph, France, 2005

[HEF00] J. Heflin et J. Hendler. Searching the web with shoe. In

Bibliographie

Artificial Intelligence for Web Search. Papers from the AAAI Workshop., 2000. 98, 99

[HER05]. N. Hernandez et N. Aussenac-Gilles. Ontoexplo : Ontologies pour l'aide à une activité de veille ou d'exploration d'un domaine. In Foix, editor, *VIème Journées de l'innovation*, Janvier 2005.

[HKR00] Herlocker J.-L., Konstan J.-A., Riedl J., Explaining Collaborative Filtering Recommendations, *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW'00)*, Pennsylvania, USA, 2000, p. 241-250.

[JH 1993] C. Jacquemin et P. Zweigenbaum. Traitement automatique des langues pour l'accès au contenu des documents. In *Le document Multimédia en Sciences du Traitement de l'Information*,

[JIA97] J. Jiang et D. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings on International Conference on Research in Computational Linguistics*, 1997. 109, 111,

[KAU87] Morgan Kaufmann , Collins, A.; and Smith, E., E., *Readings in cognitive science : A perspective from Psychology and Artificial Intelligence*, Publishers, INC, San Mateo,California, 1987.

[KIR04]. A. Kiryakov, B. Popov, I. Terziev, D. Manov, D. Ognyanoff, Semantic annotation, indexing, and retrieval, Journal of Web Semantics, 2(1), 2004.

[KOB01] Kobodner, J., L.; Hmelo, C.,E., and Narayanan, N., H., "Problem-Based Learning Meets Case-Based Reasoning", In *Proceedings of the Second International Conference on the Learning Sciences*. Charlottesville, Va.: AACE Press, 188-95.01.

[LAN95] Lang K., NewsWeeder: Learning to Filter Netnews, *Proceedings of the 12th International Conference on Machine Learning (ICML'95)*, CA, USA, 1995, p. 331-339.

[LEA94]. Leacock, C., Miller, G. A., and Chodorow, M. 1994. Using corpus statistics and WordNet relations for sense identification. Comput. Linguist. 24, 1 (Mar. 1994), 147-165.

[LEA98] C. Leacock, M. Chodorow, Combining local context and Wordnet similarity for word sense identification, in WordNet: an electronic lexical database, C. Felbaum (Ed), Cambridge, MA, The MIT Press, pp 265-283, 1998.

[LEC94] Lech, T., Wienhofen, L. AmbieAgents: A Scalable Infrastructure for Mobile and Context-Aware Information Services. In: Aarts, H., Westra, J. (eds.): Proceedings of the 4th International Conference on Autonomous Agent and Multi-Agent Systems (AAMAS 2005) (Utrecht, Netherlands, July 25-29, 94), ACM Press, New York, pp. 625-631.

[LEE05] LEE K.H., CHOY Y.C., CHO S.B. Geometric structure analysis of document images : a knowledge-based approach, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, Vol.22 No.11, pp. 1224-40.

[LER01] K. Lerman, C. Knoblock, et S. Minton. Automatic data extraction from lists and tables in web sources. In *IJCAI-2001 Workshop on Adaptive Text Extraction and Mining*, 2001.

[LIE95] [Lieberman H., Letizia: An agent that assists web browsing, *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, Canada, 1995, p. 924-929.

[LIN98] D. Lin. An information-theoretic definition of similarity. In *Proceedings of 15th International Conference On Machine Learning*, 1998.

[LYN05] Lynda Lechani tamine , Mohand Boughanem ,Accès personnalisé a l'information Approches et techniques – IRIT 2005

[MAL95] MALTZ, D. EHRLICH, E. 1995. Pointing the way: Active collaborative filtering. In CHI'95 Human Factors in Computing Systems

Bibliographie

[MAR95] Philippe Martin. *Using the WordNet Concept Catalog and a Relation Hierarchy for Knowledge Acquisition*. Proc. of Peirce'95, 4th, International Workshop on Peirce, University of California, Santa Cruz, USA, pp. 36-47, August 18th 1995.

[MAR98] Marchand, S. « Evaluation de la production scientifique de UCBL/INSA ». Mémoire du DESS Informatique Documentaire, Université Lyon 1-Enssib, 1998.

[MAT07] MATAOUI M'hamed - Reformulation de requêtes dans les systèmes de recherche d'information dans des documents XML - mgister 2007 univ BOUMERDES.

[MEK94]

[MEL02]. Melville, P.; Mooney, R.J.; Nagarajan, R.: Content-boosted collaborative filtering for improved recommendations. Proceedings of the 18th National Conference on Artificial Intelligence, 2002.: pp. 187-192.

[MIC99] MICHEL, Christine. - évaluation de systèmes de recherche d'information, comportant une fonctionnalité de filtrage, par des mesures endogène : réalisation et évaluation d'un prototype de système de recherche d'information avec filtre selon les profils des utilisateurs. – Thèse de doctorat de l'université Lyon1, Janvier 1999

[MOB04]. MOBASHER, B., JIN, X., AND ZHOU, Y. 2004. Semantically enhanced collaborative filtering on the web.

[NOU04] NOUALI omar " filtrage d'information textuelle sur les réseaux une approche hybride" thèse de doctorat USTHB-2004

[NOY01] Noy, Natalya, F.; and McGuinness, Deborah, L., "Ontology Development 101: A Guide to Creating Your First Ontology", *Stanford Knowledge Systems Laboratory Technical Report KSL-01-05* and *Stanford Medical Informatics Technical Report SMI-2001-0880*, March 2001.

[PAL97] Palopoli, L., Sacca, D. et Ursino, D. Semi-automatic, semantic discovery of properties from database schemas. In: Proc Int. Database Engineering and Applications Symp. (IDEAS), IEEE Comput, pp. 244–253, 1997.

[PAT06] Patel-Schneider, P.F., Hayes, P. et Horrocks, I. OWL Web Ontology Language Semantics and Abstract Syntax. W3C Recommendation 10 February 2004.
http://www.w3.org/TR/owl-semantics/, February 2006.

[PIN01] Pinto, H.S. et Martins, J.P. A Methodology for Ontology Integration. In Proceedings of the First International Conference on Knowledge Capture , ACM Press, 2001.

[RAD89] R. Rada, H. Mili, E. Bicknell, et M. Blettner. Development and application of a metric on semantic nets. *IEEE Transaction on Systems, Man, and Cybernetics*, 19(1) :17–30, 1989.

[RES95] Resnik, Using information content to evaluate similarity in a taxonomy, In Proceedings of the 14th joint conference in Artificial Intelligence, 1995.

[RES98] P. Resnik, Semantic similarity in a taxonomy: an information based measure and its application to problems of ambiguity in natural langage, Journal of Artificial Intelligence Research, volume 11, pp 95-130, 1998

[RIS79] VAN RIJSBERGEN C.J. Information Retrieval. London, Boston : Ed. Butterworths, 1979, 208 p. ISBN 0408709294

[RUC97] Ruch P., Wagner J., Bouillon P., Baud R.H., Rassinoux A.M., Scherrer J.R., 1997, *MEDTAG: tag-like semantics for medical document indexing*, Proc AMIA Symp., p. 137-41.

[SAL70] G. Salton. The SMART retrieval system : Experiments in automatic document processing. Prentice Hall. 1970.

Bibliographie

[SAL71] G. Salton, The Smart Retrieval System, Prentice Hall, Englewood Cliffs, NJ, 1971

[SAL83] Salton, G., Fox, E., and Wu, H. Extended boolean information retrieval. Communications of the ACM, 26(11) :1022–1036. 1983.

[Sal84] G. Salton and M. McGill. Introduction to modern information retrieval. McGraw-Hill Int. Book Co. 1984.

[SAL88] Salton G., Buckley C., Term Weighting Approaches in Automatic Text Retrieval, *Information Processing and Management*, vol. 24 (5), 1988, p.513-523.

[SAN94] M. Sanderson. 1994. Word sense disambiguation and information retrieval. In Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pages 142-151, Springer-Verlag.

[SAR04l]. Sarwar, B. M., Karypis, G., Konstan, J. A. & Riedl, J. (2004), Item-based collaborative filtering recommendation algorithms.

[STA01] S. Staab et A. Maedche. Knowledge portals - ontologies at work. *AI Magazine*, 2001.

[SUR 02] Sure, Y.; Erdmann, M.; Angele, J.; Staab, S.; Studer, R.; and Wenke, D., "OntoEdit: Collaborative Ontology Engineering for the Semantic Web", In *Proceedings of the International Semantic Web Conference 2002 (ISWC 2002)* , Sardinia, Italia, June 9- 12 2002.

[TER93] D.B Terry. A tour through tapestry. In *COOCS*, pages 21-30. ACM,1993.

[TMA01] Mohamed.Tmar. *Modèle auto-adaptatif de filtrage d'information: apprentissage*
[TMA02] *incrèmental du profil et de la fonction de décision*. Université Paul Sabatier, Toulouse, 2002.

[VAS99] Vassileva, J. « A Task Centered Approach for User Modelling in Hypermedia Office Documentation System », *User Modelling and User-Adapted Intercation,* vol 6, n°2-3, p. 185-224, 1999.

LOA96]. Van Loan. G. Golub et C. Matrix computations. The Johns Hopkins University Press, Baltimore, 1996.

ANNEXE A : Outil Matlab

.'	transposition
:	conversion matrice en vecteur
Axis	Paramètres des axes
break	interrompt une boucle for ou while
cd	change le directory courant
clear	efface les variables et fonctions de la mémoire
corrcoef	coefficients de corrélation
cov	matrice de covariance
cputime	temps CPU écoulé
delete	suppression de fichiers
Det	Déterminant d'une matrice
diag	matrice diagonale
diag	création ou extraction de la diagonale
diag	création ou extraction de la diagonale
disp	affiche une matrice de texte
eig	valeurs et vecteurs propres
else	complète if
elseif	complète if
end	terminaison de if, for et while
error	affiche un message et interrompt l'exécution
etime	durée d'exécution
eye	matrice identité
find	Indices non nulles
fliplr	retournement gauche-droit
flipud	retournement haut-bas
for	instruction de répétition avec compteur
Format	Le format de sortie
getenv	renvoie la variable d'environnement
Grid	Lignes de textures
help	aide
if	test conditionnel
inv	inversion

length	renvoie la longueur d'un vecteur
ls	liste les fichiers
lscov	moindres carrés avec covariance connue
matlabrc	M_file principal de lancement
max	valeur max d'un vecteur
mean	valeur moyenne d'un vecteur
median	valeur médiane d'un vecteur
min	valeur min d'un vecteur
ones	matrice de 1
orth	orthogonalisation
path	Défini les chemins d'accès aux fichiers et fonctions
Plot	Graphique 2D
prod	produit des éléments d'un tableau
pwd	affiche le directory courant
rand	nombres aléatoires à répartition uniforme
reshape	redimensionnent
return	retour
rot90	rotation de 90°
size	renvoie la taille d'une matrice
sort	tri en ordre croissant
startup	M_file de lancement de MatLab
std	écart type
sum	somme des éléments d'un tableau
svd	décomposition en valeurs singulières
tic, toc	affiche le début et la fin d'exécution d'un processus
Title	Titre d'un graphe
trace	Trace d'une matrice
tril	partie triangulaire inférieure
triu	partie triangulaire supérieure
while	instruction de répétition avec test
xlabel	Titre de l'axe x
zeros	matrice de 0

Annexe A MATLAB

Editor - E:\mat3\AnnexeMATLAB.m*

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% La matrice d'evaluation des ITEMS pat les USERS      %%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


E=[5    3   4   3   3   5   4   1   5   3   2   5   5   5   5   5   3   4   5   4   1   4   4   3   4   3   2
4   0   0   0   0   0   0   0   0   2   0   0   4   4   0   0   0   0   3   0   0   0   0   0   4   0   0   0
0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
0   0   0   0   0   0   0   0   0   0   4   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
4   3   0   0   0   0   0   0   0   0   0   0   0   0   0   0   4   0   0   0   3   0   0   4   3   0   0   0
4   0   0   0   0   0   2   4   4   0   0   4   2   5   3   0   0   0   4   0   3   3   4   0   0   0   0   2
0   0   0   5   0   0   5   5   5   4   3   5   0   0   0   0   0   0   0   0   0   5   3   0   3   0   4   5
0   0   0   0   0   0   3   0   0   0   3   0   0   0   0   0   0   0   0   0   0   5   0   0   0   0   0   0
0   0   0   0   0   5   4   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
4   0   0   4   0   0   4   0   4   0   4   5   3   0   0   4   0   0   0   0   0   5   5   0   0   0   0   0
0   0   0   0   0   0   0   4   5   0   2   2   0   0   5   0   0   0   0   0   0   4   0
.......
.......

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%% La matrive semantique des ITEMS  ITEMS & attributs      %%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

 ES=[1  1995    0   0   0   1   1   1   0   0   0   0   0   0   0   0   0   0   0   0   0;
2    1995    0   1   1   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0;
3    1995    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0;
4    1995    0   1   0   0   0   1   0   0   1   0   0   0   0   0   0   0   0   0   0;
5    1995    0   0   0   0   0   0   1   0   1   0   0   0   0   0   0   0   1   0   0;
6    1995    0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0;
7    1995    0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   1   0   0   0;
8    1995    0   0   0   0   1   1   0   0   1   0   0   0   0   0   0   0   0   0   0;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%        La similarite par evaluation                      %%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
[lig,col]=size(E);
S=eye(col);
for j=1:col-1
    v1=E(:,j);
    for k=j+1:col
        v2=E(:,k);
        inc1=find(v1 & v2);
        x=(sum(v1(inc1).*v2(inc1)))/(sqrt(sum(v1(inc1).^2)*sum(v2(inc1).^2)));
        S(j,k)=x;
        S(k,j)=x;
    end ;
end;
%disp('la matrice devaluation est: ');
E ;
%disp('la matrice de similarite resultante est : ');
S;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%        La Prediction par evaluation                      %%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

P=E-E;
ii=1;
vp2=randnc(1,col-10);
for nbv=10:10:col-10
    vp1=[10:10:col-10];
```

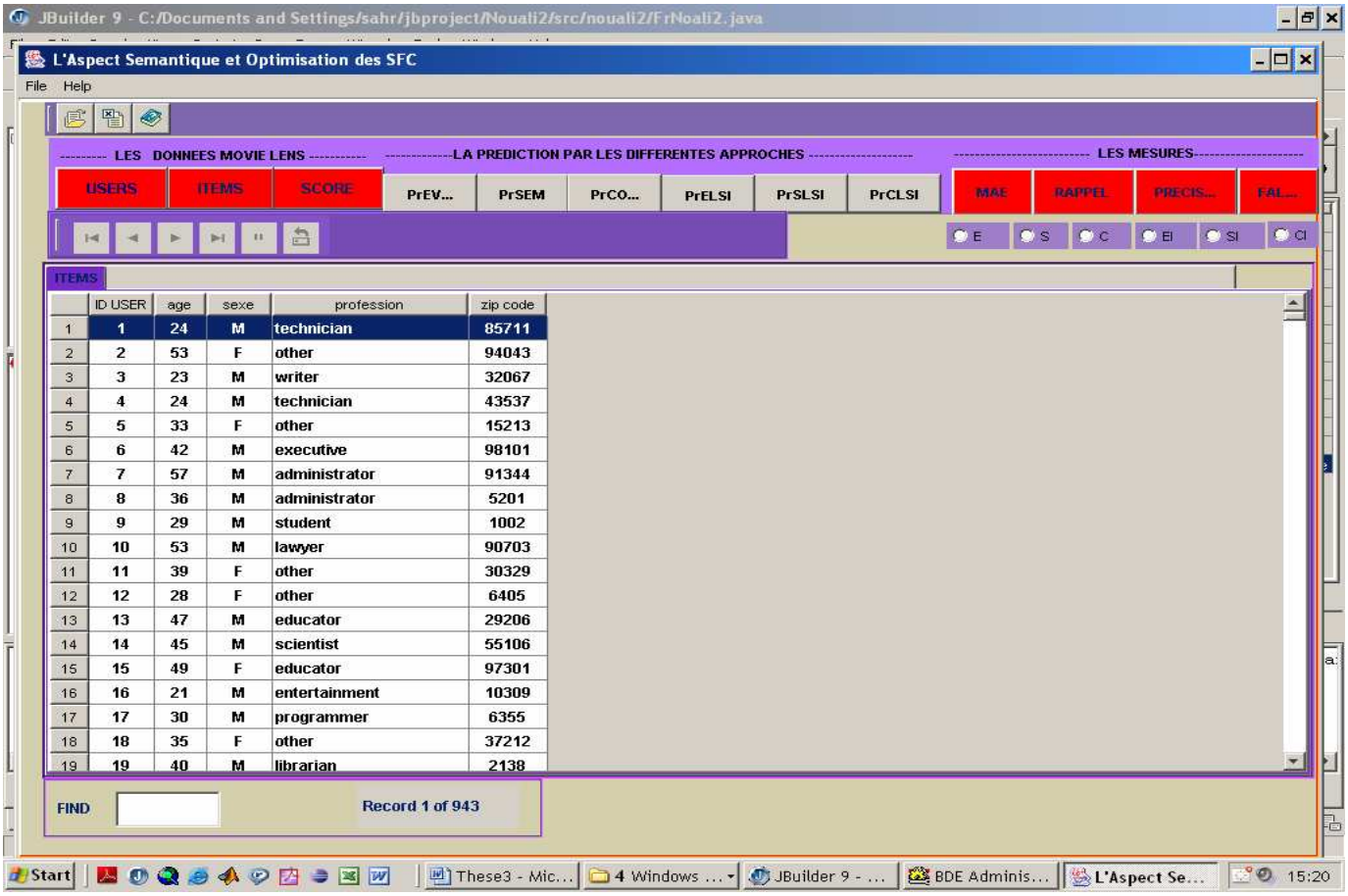
Annexe A MATLAB

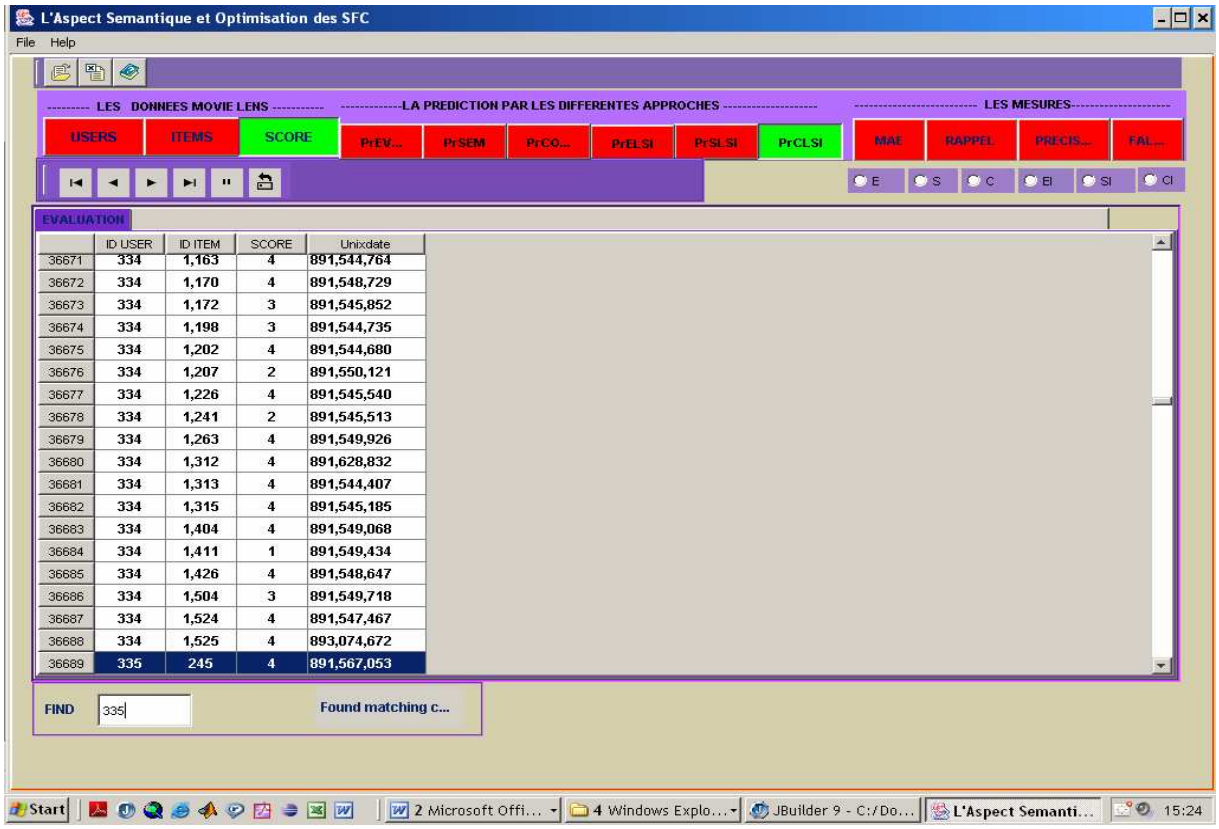
```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%       similarite semantique;                            %%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%


SS=eye(col1);
for j=1:col1-1
    v1=ES(:,j);
    for k=j+1:col1
        v2=ES(:,k);
        inc1=find(v1 & v2);
        x=(sum(v1(inc1).*v2(inc1)))/(sqrt(sum(v1(inc1).^2)*sum(v2(inc1).^2)));
        SS(j,k)=x;
        SS(k,j)=x;
    end ;
end;
%disp('la matrice devaluation semantique est: ');
ES
SC=(S+SS)/2;   %similarite combinee
PC=E-E;        % prediction combinee
ii=1;
vp2=randnc(1,col-10);
for nbv=10:10:col-10
    vp1=[10:10:col-10];
for i=1:lig
    for j=1:col
        var1=0;
        var2=0;
        SC(j,j)=0;
        vec1=SC(:,j);
        for k=1:nbv
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%  la decomposition et la reduction de la matrice evaluation E SVD-10     %%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
E=E(1:100,1:200);
[u s v]=svds(E,10);
E=u*s*v';
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%        Le vecteur d'erreur MAE            %%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

mae=E(:,1)-E(:,1);
for i=1:lig
    some=0;
    for j=1:col
        some=some+abs(E(i,j)-P(i,j));
    end
    mae(i)=some/col;
end;
%disp('le vecteur des erreurs moyens est:');
mae;
vp2(ii)=mean(mae);
ii=ii+1;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%        l'affichage du resultat            %%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
cmp=1:1:length(vp1);
x=vp1(cmp);
y=vp2(cmp);
%subplot(221);
plot(x,y,'-r','LineWidth',2,'MarkerEdgeColor','r');
title('LE MAE EVALUATION (base items-SVD10)');
legend('base items','r','location','northwest');
xlabel('K NOMBRE DE VOISINS'),ylabel('M  A  E'), grid
```

Annexe A MATLAB

ANNEXE B : L'interface Graphique





Annexe B L'interface graphique