UQÀM
Université du Québec à Montréal

DÉPARTEMENT D'INFORMATIQUE

DOCTORAT EN INFORMATIQUE COGNITIVE

PRÉSENTATION DU PROJET DE RECHERCHE

Modèle multi-agent pour le filtrage collaboratif de l'information.

Nom de l'étudiant : Zied Zaïer.

Nom des directeurs de recherche : Robert Godin.

Luc Faucher.

SESSION D'AUTOMNE 2003

Tables des matières

Liste des figures

Liste des tableaux

1. Introduction

Le réseau Internet connaît depuis quelques années un accroissement très important du nombre d'utilisateurs, du nombre d'ordinateurs connectés et de la quantité d'informations qui y est disponible. La nature même de ce réseau en a fait, dès le départ, un outil potentiel pour les documentalistes. Les avantages du réseau Internet sont en effet nombreux : il est relativement bon marché, il offre une interface de navigation simple et unifiée, il est un moyen de communication rapide et efficace et surtout un lieu de publication « dynamique », à savoir que quiconque peut y publier des documents et les remettre à jour aisément et rapidement.

Tout chercheur peut ainsi rendre accessibles ses travaux sans les contraintes de l'édition papier (Prax, 1998). De même, la communication scientifique est facilitée par le courrier électronique et les innombrables forums et « lettres » d'information en version électronique. Les possibilités d'accès à la connaissance sont donc potentiellement plus rapides grâce au réseau Internet. Par ailleurs, l'accès à l'information par l'intermédiaire du réseau n'est plus limité aux informations publiques et gratuites: les banques de données commerciales proposent peu à peu des accès à leurs sources d'informations via le réseau. De même, des journaux proposent le contenu de leurs archives gratuitement ou de manière payante.

Le réseau est donc en train de s'imposer comme source d'informations complémentaire des services classiques. Il ne suffit pas d'être au sommet d'une montagne d'ouvrages pour devenir soudain plus intelligent. Il faut pouvoir retrouver l'information pertinente sur le réseau et savoir l'exploiter. Or, la nature même de l'Internet veut qu'elle soit éparpillée, non centralisée.

La liberté complète de publication sur le web a pour corollaire une grande hétérogénéité de son contenu : sites web non remis à jour, fausses informations, publicités… Cette excessive liberté a pour inconvénient l'apparition de nombreuses sources d'informations non pertinentes pour le documentaliste. L'édition papier possède l'avantage d'être soumise à un comité éditorial qui juge de la qualité du document à éditer. Rien de tel avec l'Internet. Cela fait sa force, mais aussi sa faiblesse. La structure même du réseau rend difficile la recherche de l'information. Structure très décentralisée, elle est organisée de manière radicalement différente des bases de données ordinaires.

Ces contraintes ont transformé la recherche d'informations sur Internet en une démarche ardue. La quête des ressources pertinentes sur un sujet donné est donc logiquement devenue un problème essentiel des internautes. Certains des sites les plus visités sont ceux qui ont pour objectif d'aider les utilisateurs dans leurs recherches au moyen de répertoires (Yahoo, Magellan…) ou au moyen de

moteurs de recherche ayant pour ambition d'indexer le texte intégral de la totalité des pages web (Google, Altavista, Hotbot…). Mais devant l'extraordinaire croissance et diversité des ressources disponibles sur le réseau, ces outils se révèlent de plus en plus insuffisants, autant pour les internautes que pour les documentalistes. Les uns comme les autres se trouvent confrontés à un très grand nombre de réponses retournées par les moteurs de recherche et au problème d'un ensemble de ressources à la fois "dynamique" (car fréquemment remis à jour) et non centralisé (Andrieu, 1996).

Une évolution des moteurs de recherche et autres annuaires est donc nécessaire. Des outils se développent peu à peu qui ont pour objectif d'automatiser certaines des tâches remplies actuellement par les internautes en quête d'informations sur le réseau. Ces nouveaux outils tendent vers un idéal qui est celui d'un agent logiciel "intelligent" qui assisterait l'internaute dans ses recherches, les rendant à la fois plus complètes et plus rapides.

Entre les moteurs de recherche, les annuaires, les métamoteurs d'une part et les agents intelligents d'autre part, s'est donc développée toute une gamme de logiciels et de services intermédiaires, à mi-chemin entre l'une et l'autre de ces catégories, empruntant certaines caractéristiques aux uns et aux autres. Leur objectif est le même: rendre plus rapides et plus efficaces les recherches des usagers de l'Internet.

2. Présentation du projet de recherche

2.1. Problématique

L'utilisation d'agents se justifie par les caractéristiques spécifiques de l'Internet. La structure et l'organisation du réseau mondial expliquent à la fois les raisons qui nécessitent leur emploi et les contraintes qui limitent leur utilisation. Nous verrons à quelles contraintes structurelles se heurtent les systèmes de recherche d'information.

2.1.1. Problèmes issus d'Internet

Pour comprendre l'intérêt d'outils de plus en plus sophistiqués pour la recherche d'informations sur Internet, nous commencerons par rappeler quelques-unes des caractéristiques de ce réseau qui expliquent les insuffisances des outils de recherche existants à ce jour.

- Un espace éditorial ouvert

Sur le réseau Internet, chacun peut devenir producteur d'informations. Un ordinateur, un modem, un fournisseur d'accès, quelques notions d' HTML et il est possible de créer son propre site web. Cela signifie que les intermédiaires traditionnels (en premier lieu les éditeurs) sont en partie court-circuités par cette relation directe établie entre le lecteur et l'auteur. Chaque université, chaque centre de recherche, chaque entreprise peut devenir en quelque sorte son propre éditeur.

Il est clair que cette liberté ne présente guère d'intérêt pour les professionnels de la recherche d'informations lorsqu'elle est utilisée pour mettre en ligne des chansons ou des photos de famille. Elle est pertinente, en revanche, quand il s'agit d'être au courant de l'actualité des centres de recherches, des universités, des entreprises… Cette liberté éditoriale explique le grand dynamisme du réseau Internet. L'information peut se renouveler aisément et rapidement. Les informations scientifiques et techniques circulent de manière très rapide. Les échanges et les collectes d'informations se font beaucoup plus facilement. La diffusion des connaissances dépend ainsi, dans une moindre mesure qu'auparavant, du rythme des revues et de ses contraintes (coûts, pagination…) (Chartron, 1996).

D'une telle situation, il résulte une certaine anarchie éditoriale. Les sources présentes sur le web sont d'une grande variété. La page web de l'étudiant côtoie le serveur du MIT (*Massachusetts Institute of Technology*); les sources payantes sont concurrencées par des sources gratuites. La fonction de validation de l'information par un comité éditorial n'est plus assurée. Le bon peut donc côtoyer le pire et l'information pertinente risque d'être masquée par la publicité et la désinformation, qu'elle soit volontaire ou non. Il y a donc un besoin d'identification des ressources réellement pertinentes. Certains services se consacrent à cette tâche en commentant les sites qu'ils répertorient (*Magellan*, *Nomade*…). Autre problème : les ressources apparaissent et disparaissent sans que la plupart des utilisateurs en soient informés. Trouver les ressources existantes et surtout les plus récentes et les plus fiables est une véritable gageure.

- L'information disponible sur le web est dynamique

Le web est affranchi des rythmes de périodicité de parution. Cela permet, si nécessaire, une mise à jour quotidienne, à l'inverse cependant un site peut rester inchangé pendant des mois voire des années ou changer de localisation ou d'appellation. L'enjeu majeur entraîné par cette situation est de savoir quels sites sont fréquemment renouvelés et quand l'ont-t-ils été pour la dernière fois (Lawrence et Giles, 1999).

- La structuration des informations offertes sur le réseau est faible

Les informations disponibles sur Internet sont d'une nature hétérogène. Différents types et formats de documents se côtoient, rendant plus ou moins aisée la recherche d'informations.

Les données proposées sur le web sont en général peu structurées malgré le développement des champs métadonnées dans le langage HTML. Des indications, appelées balises, sont censées permettre une indexation par nom, sujet, etc. du document HTML auquel elles se rapportent. Certains services web tiennent compte de ces balises pour réaliser un meilleur calcul de pertinence des documents. Une telle « indexation » trouve malheureusement rapidement ses limites dans la mesure où elle est totalement libre, non contrôlée, non obligatoire et souvent générée automatiquement par les logiciels d'édition HTML.

Cette faible structuration explique la prédominance des moteurs de recherche existants, indexant le texte intégral des documents HTML.

2.1.2. Caractéristiques et insuffisances des systèmes existants de recherche d'informations

Ensemble de plusieurs dizaines de millions de documents, le web connaît plusieurs modes d'indexation. Aucun n'est complet et totalement pertinent (Chartron, 1996 ; Lardy, 1996a ; Lardy, 1996b).

- Les répertoires

Ils sont organisés autour d'une classification intellectuelle. L'exemple le plus célèbre est le service « *Yahoo* ». Les sites répertoriés dans ce type de services sont classés par thèmes, sous-thèmes, etc. Les catégories utilisées ne sont pas celles employées généralement dans les bibliothèques généralistes, mais elles s'en approchent. Pour "*Yahoo"* les catégories de classification sont les suivantes : "Arts and Humanities", "Business and Economy", "Computers and Internet", "Education", "Entertainment", "Government", "Health", "News and Media", "Recreation and Sports", "Reference", "Regional", "Science", "Social Science", "Society and Culture".

Ce type de classification a l'avantage de la proximité avec les modes traditionnels de classification. La recherche peut se faire de deux manières : soit en choisissant le thème puis le sous-thème, etc. soit en faisant une recherche par mots, sur les titres des thèmes et sous-thèmes de l'annuaire. Dans tous les cas, il faut avoir une idée du thème dans lequel peut être classée l'information recherchée. Il ne faut

pas non plus s'attendre à trouver instantanément la réponse à sa question. Les répertoires donnent des références de sites et non des adresses de pages. Ils donnent en général une liste de sites. À l'utilisateur de les visiter et de voir s'ils conviennent ou non à sa demande ou non. Les répertoires sont augmentés soit par inscription de leurs sites par leurs auteurs, soit par des moteurs de recherche capables d'assurer une classification des pages recueillies (une vérification humaine est nécessaire).

Les répertoires ont pour défauts :

- ✓ une absence d'exhaustivité (augmentation de la base répertoriée par inscription des auteurs),
- ✓ des insuffisances en cas de recherche thématique très précise (nom de produits, de personnes…).

Précurseurs dans le domaine de la recherche et de l'indexation de l'information sur le Web, les répertoires ont donc été rapidement complétés par des "moteurs de recherche" (Lawrence et Giles, 1999.).

➢ Les moteurs de recherche

Ils ont pour nom « Google », « *Altavista* », « *Excite* », « *HotBot* », etc. Leur ambition est de réaliser une indexation automatique et complète de l'ensemble des pages web existantes. Ces logiciels présentent souvent les défauts et les qualités inverses des répertoires. Ambitionnant l'exhaustivité, ces moteurs indexent la totalité des termes contenus dans les pages web. Un « *robot* » (*spider*) parcourt en permanence les pages web du monde entier. Il enregistre et indexe, selon la méthode choisie, la totalité ou une partie des termes de chaque page dans la base de données du moteur en question. En effet, généralement, une étape préliminaire de normalisation du texte est nécessaire, c'est-à-dire se débarrasser des caractères spéciaux, puis utiliser une liste de mots vides pour supprimer tous les mots qui ne sont pas porteurs de sens (pronoms, prépositions, conjonctions, articles, etc.). Puis, sur les mots qui restent, on effectue une troncature, c'est-à-dire une analyse des formes morphologiques des mots, pour ne garder que leur racine. Lorsque l'utilisateur formule une demande, le moteur recherche dans sa base de données, le ou les termes recherchés, et liste les références des pages où ont été repérés, au moins une fois, le ou les termes recherchés.

Une telle indexation, pour séduisante qu'elle soit, est limitée. Certes, il est quasiment impossible de ne pas trouver une réponse à une question. Mais reste à trouver la bonne. La simple présence du terme recherché dans un document suffit en effet, dans la logique de ces moteurs, à le rendre potentiellement pertinent. Certes, des calculs de pertinence sont effectués par les moteurs de recherche. Cependant, un terme peut être employé dans de multiples contextes et pour de nombreuses raisons sans que le

document qui le contient ne soit pour autant pertinent pour l'utilisateur. Conséquence de ce mécanisme : l'exhaustivité relative des moteurs est contrebalancée par une surabondance de documents non pertinents parmi les références des pages web (URL) qu'ils indexent.

Ajoutons à cela que même les principaux moteurs de recherche ne recensent pas l'intégralité du web. En effet, même les meilleurs d'entre eux ne peuvent assurer une mise à jour quotidienne de leurs informations. Ces moteurs de recherche n'interrogent pas en effet directement le web, mais leur base de données contenant les termes décrivant chaque page web. Or, la remise à jour de cette indexation peut prendre un certain temps, des semaines, voire des mois. Pour ces diverses raisons, les moteurs de recherche ne peuvent assurer une indexation totalement fiable d'un univers dynamique et changeant comme celui du web. Les moteurs de recherche peuvent difficilement suivre les mises à jour de sites où l'information change quotidiennement. Ce sont des outils à la fois indispensables et insuffisants (Lawrence et Giles, 1999).

2.2. Définition de la recherche

Dans le cadre de cette recherche exploratoire, la démarche méthodologique suivie a été basée sur le cadre de Basili (Abran, Framboise et Bourque, 1999), adapté pour son application aux cas des études de type exploratoire (Annexe 1). La définition de la recherche à partir de ce cadre, a permis d'établir la motivation, les questions de recherche, les objectifs de recherche et les utilisateurs de cette recherche (pour plus de détails, voir la section 5).

2.2.1. Motivation

Tous les problèmes évoqués dans la partie précédente sont essentiellement dus à l'absence, à l'heure actuelle, d'outils performants. Parmi les solutions proposées sur le marché, nous pouvons distinguer :

- *Les agents sociables* : Les agents sociables sont les applications que nous pouvons le plus facilement rapprocher des futurs agents intelligents. En effet, les agents sociables apprennent les goûts des utilisateurs en même temps qu'ils leur fournissent des documents. Les utilisateurs évaluent les documents fournis, ce qui permet d'améliorer la qualité des recherches ultérieures. Les deux produits leaders sont Wisewire et Firefly (Stanley, 1997).
- *Les agents semi-intelligents :* ces produits sont considérés comme semi-intelligents, car ils leur manquent, dans la plupart des cas, de nombreuses caractéristiques des véritables

agents intelligents. En premier lieu la capacité à réagir à leur environnement et surtout à échanger avec les autres agents (Piechaczyk, 1996).

- *Les agents de navigation en local (off Line) :* Les agents de navigation en local permettent de télécharger sur le disque dur vos sites Web préférés et de les consulter en local. Ils permettent d'autre part de surveiller la remise à jour régulière de ces sites, c'est-à-dire que périodiquement le logiciel va vérifier si le site a été modifié. Certains logiciels permettent par ailleurs de faire des recherches au sein des pages téléchargées sur le disque dur (Derudet, 1997).
- *Les agents guides* : Ce sont des applications qui ont pour objectif d'accompagner et d'assister les utilisateurs dans leur navigation par une série de fonctionnalités variées. Les produits les plus connus sont WBI, Webtamer et Alexa (Haskin, 1997).
- *La webdiffusion ("push media" ou "webcasting") :* Le principe de base de la webdiffusion est de fournir des informations pertinentes à l'utilisateur de manière automatique, sans que celui-ci n'ait à les rechercher sur Internet (Renaud Chavanne, 1997).

La motivation principale de ce travail est d'essayer d'*améliorer* les outils d'accès à l'information sur Internet existants sur le marché. Plus précisément, nous allons approfondir les points suivants :

- Filtrage de l'information
- Plate-forme d'agents adaptables.
- Collaboration entre agents.
- Adaptation de divers profils utilisateur basée sur apprentissage.

2.2.2. Question de recherche

La question principale à laquelle nous allons essayer de répondre est la suivante :

Est-il possible d'obtenir un *modèle* coopératif, basé sur une architecture Peer-To-Peer, capable de comportements sociaux complexes? Cette question peut être subdivisée en trois sous-questions :

- Quelles sont les caractéristiques des modèles qui existent sur le marché ?
- Quels sont les points forts et les faiblesses de chaque modèle ?
- Quelles sont les améliorations possibles ?

2.2.3. Objectifs de recherche

L'objectif principal de cette recherche est de développer et d'*évaluer* un modèle, basé sur une architecture Peer-To-Peer, pour le repérage de l'information collaboratif exploitable dans un contexte Internet. À cette fin, nous allons utiliser la méthodologie suivante :

- ✓ Faire une étude de l'existant.
- ✓ Proposer un modèle.
- ✓ Implémenter le prototype.
- ✓ Faire une évaluation du prototype.

2.2.4. Utilisateurs de la recherche

Le résultat final de cette étude est un prototype fonctionnel qui peut être utilisé par des internautes ou des professionnels de la recherche d'informations.

2.3. Étude de l'existant

Nous nous proposons de fournir, ici, quelques pistes de recherche pour répondre à notre problématique. En effet, lors de notre revue de la littérature, nous avons pu identifier les travaux les plus pertinents pour notre sujet de recherche. Étant donné que le sujet choisi est tout à fait interdisciplinaire, les documents étudiés viennent de différents champs de recherche, à savoir le filtrage l'information, les architectures distribuées, les systèmes multi-agents.

2.3.1. Filtrage de l'information

Généralement, l'accès à l'information sur Internet se fait selon deux méthodes. D'une part, une recherche active de documents est effectuée via des systèmes de recherche d'informations. Selon Belkin, ces derniers ont pour fonction « d'amener à l'utilisateur les documents qui vont lui permettre de satisfaire son besoin en information » (Belkin et Croft, 1992). D'autre part, la réception de documents jugés intéressants par des systèmes de filtrage d'information. En opposition avec les outils de recherche d'informations, le filtrage d'information ne nécessite pas une formulation systématique du besoin de l'utilisateur. Ainsi, cela procure à l'utilisateur une économie d'effort, mais également une certaine sérénité. D'après Croft, le système de filtrage d'informations « achemine des documents qui se

présentent vers des groupes de personnes, en se basant sur leurs profils à long terme élaborés à partir de donner d'apprentissage » (Croft, 1993).

Pour résumer, le filtrage d'informations est souvent défini comme l'élimination de données indésirables sur un flux entrant, plutôt que la recherche de données spécifiques sur ce flux. Nous vous présentons ci-dessous un tableau comparatif détaillé des différences entre la recherche d'informations et le filtrage d'informations.

	Recherche d'informations	Filtrage d'information
Approche	Trouver l'information recherchée	Filtrer l'information non désirée
Livraison	Corpus statique, sur demande	Flux dynamique
Persistance	Des besoins à court terme	Des intérêts à long terme
Analyse du contenu	Utilise des mots-clés	Pour le filtrage basé sur le contenu Pas pour le filtrage collaboratif.
Fonctionnalités	Non personnalisé, non adaptatif, non dynamique, à court terme.	Personnalisé, s'attaquent aux changements du profil de l'utilisateur, filtre dynamiquement d'informations entrantes, à long terme.

Tableau 1. Recherche d'informations versus filtrage d'informations.
(Belkin et Croft, 1992)

La figure ci-dessous nous présente un modèle type de filtrage d'information. Comme nous l'avons vu plus haut, les systèmes de filtrage présupposent l'existence d'un flux de données entrant qui est émis par une source distante ou envoyé directement par d'autres sources (d'autres utilisateurs). En ce qui concerne le filtrage, il est basé sur des descriptions d'utilisateurs ou/et de groupes d'utilisateurs constituant des profils. Ces derniers symbolisent, dans la majorité des cas, une compilation de domaines d'intérêt à long terme. Enfin, les documents résultants du filtrage sont évalués pour mettre à jour les profils et les thèmes d'intérêt.

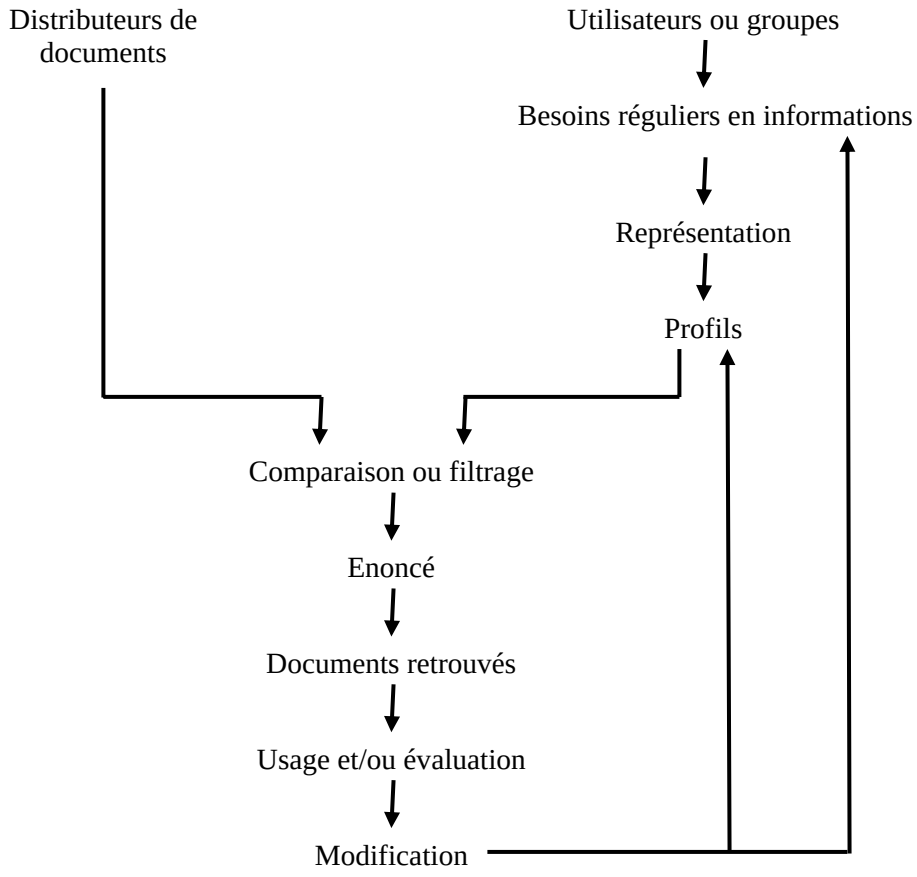

Figure 1. Modèle général pour le filtrage d'information,
inspiré de (Belkin et Croft, 1992).

Pour terminer, une revue de la littérature nous montre qu'il existe trois grandes familles de systèmes de filtrage d'information :

- Le filtrage basé sur le contenu (aussi appelé filtrage cognitif) : le choix des documents proposés est basé sur une comparaison des thèmes abordés dans les documents par rapport aux thèmes intéressant l'utilisateur ;
- le filtrage collaboratif : la sélection des documents est basée sur des appréciations d'utilisateurs jugés semblables sur ses documents ;
- le filtrage hybride : combine les deux approches précédentes.

2.3.1.1 Le filtrage basé sur le contenu

L'idée principale est d'utiliser le contenu des documents lus par les utilisateurs pour aider à la caractérisation de ceux-ci. En effet, le filtrage basé sur le contenu s'appuie sur un profil qui décrit le besoin de l'utilisateur du point de vue thématique, de façon analogue à une requête qui serait destinée à un système de recherche d'informations. Plus précisément, le filtrage basé sur le contenu peut être vu comme un système de recherche d'informations dont la fonction de correspondance entre une requête et un corpus de documents joue le rôle d'un filtre permanent entre un profil (sorte de requête à long terme et évolutive) et le flot de documents entrant (sorte de corpus évolutif). De ce fait, deux fonctionnalités centrales ressortent, pour un système de filtrage. D'une part, la sélection des documents pertinents vis-à-vis du profil. D'autre part, la mise à jour du profil en fonction du retour de pertinence fourni par l'utilisateur sur les documents qu'il a reçus. Cette dernière se fait par intégration des thèmes abordés dans les documents jugés pertinents.

Vu la nature des données (du texte), il faut procéder à un prétraitement pour transformer le document qu'on veut analyser en une représentation convenable pour l'algorithme (Aas et Eikvil, 1999). Pour ce faire, il faut, dans un premier temps, choisir le niveau d'analyse linguistique à utiliser. En effet, différents niveaux d'analyse sont possibles. On peut décider de ne pas rentrer en profondeur dans le texte, et de ne considérer que certaines informations superficielles comme le titre ou les mots clés associés au document ; ou on peut considérer le texte comme un sac de mots (approche « bag of words ») en supposant que tous les mots sont indépendants les uns des autres; ou bien on peut considérer des n-grammes, c'est-à-dire conserver les groupes de n caractères ; enfin, on peut également faire appel à des informations syntaxiques (données par des grammaires) ou sémantiques (données par exemple par un thesaurus : dictionnaire hiérarchique décrivant des relations sémantiques entre termes), pour cibler les concepts plus généraux émanant du texte.

Ensuite, il s'agit généralement, dans un premier temps, de normaliser le texte, c'est-à-dire de se débarrasser des caractères spéciaux, puis d'utiliser une liste de mots vides pour supprimer tous les mots qui ne sont pas porteurs de sens (pronoms, prépositions, conjonctions, articles, etc.). Puis, sur les mots qui restent, on effectue une troncature, c'est-à-dire une analyse des formes morphologiques des mots, pour ne garder que leur racine. Enfin, on regroupe les mots identiques, en les comptant ou non, selon l'algorithme qu'on veut développer. Pour stocker ensuite en mémoire le document analysé, la représentation d'un document la plus utilisée est le modèle vectoriel, c'est-à-dire qu'un document est représenté par le vecteur des mots qui apparaissent le plus souvent dans le document. Selon

l'algorithme choisi, on associera à chaque mot composant le vecteur son poids dans le document aussi appelé « word frequency weighting ») (son nombre d'occurrences, ou le fait qu'il soit mis en valeur, Cosine, Okapi, TFIDF : « Term Frequency, Inverse Document Frequency », etc.), ou bien on ne s'intéressera qu'à l'absence ou la présence d'un mot (algorithme de Bayes). Une fois la classification thématique de document effectuée, on peut envisager d'intégrer les mots majoritaires à un profil d'utilisateur. La représentation la plus simple d'un profil d'utilisateur serait alors un vecteur des thèmes et mots les plus fréquents dans les lectures de celui-ci. Cependant, Widyantoro propose d'autres formes plus complexes de profil d'utilisateur (Widyantoro et al, 2000).

Pour conclure, ces systèmes présentent un certain nombre de limites. Tout d'abord, citons la difficulté d'indexation de documents multimédia. En effet, le profil utilisateur peut prendre diverses formes, mais il repose toujours sur des mots qui seront comparés aux mots qui composent le document. De ce fait, l'impossibilité d'indexer des documents multimédias. En outre, l'incapacité à traiter d'autres critères de pertinence que les critères strictement thématiques pose également un problème. En effet, le filtrage des documents basé sur le contenu ne permet pas d'intégrer d'autres facteurs de pertinence que le facteur thématique. Pourtant, il existe de nombreux autres facteurs de pertinence par exemple l'adéquation entre le public visé par l'auteur et l'utilisateur (age, position géographique, etc.), ou encore la qualité scientifique des faits présentés, la fiabilité de la source d'information, etc. En plus, l'effet dit « entonnoir » restreint le champ de vision des utilisateurs. En effet, le profil évolue toujours dans le sens d'une expression du besoin de plus en plus spécifique, qui ne laisse pas de place à des documents pourtant proches, mais dont la description thématique diffère fortement. Par exemple, lorsqu'un nouvel axe de recherche surgit dans un domaine, avec de nouveaux termes pour décrire les nouveaux concepts, ces termes n'apparaissent pas dans le profil, ce qui élimine automatiquement les documents par filtrage ; l'utilisateur n'aura donc jamais l'occasion d'exprimer un retour de pertinence positif envers ce nouvel axe de recherche. Enfin, le problème dit de la « masse critique » limite les performances du système. En effet, il faut un certain nombre de documents et d'utilisateurs pour que le système commence à donner des résultats.

2.3.1.2 Le filtrage collaboratif

Le filtrage collaboratif ou encore appelé système de recommandation (terme introduit en 1994) se base sur l'hypothèse que les gens à la recherche d'information devraient pouvoir se servir de ce que d'autres ont déjà trouvé et évalué. Cette approche résout les problèmes de l'approche basée sur le contenu sémantique ; il devient possible de traiter n'importe quelle forme de contenu. Pour ce faire, pour chaque utilisateur d'un système de filtrage collaboratif, un ensemble de proches voisins est identifié, et la décision de proposer ou non un document à un utilisateur dépendra des appréciations des membres de son voisinage (Sarwar et al, 2000).

La figure 2 présente l'organisation générale d'un système de recommandation. Typiquement, les étapes du système sont les suivantes:

- collecter les appréciations des utilisateurs sur les pages qu'ils parcourent,
- intégrer ces informations dans les profils d'utilisateur,
- utiliser ceux-ci ensuite pour aider les utilisateurs dans leurs prochaines recherches.

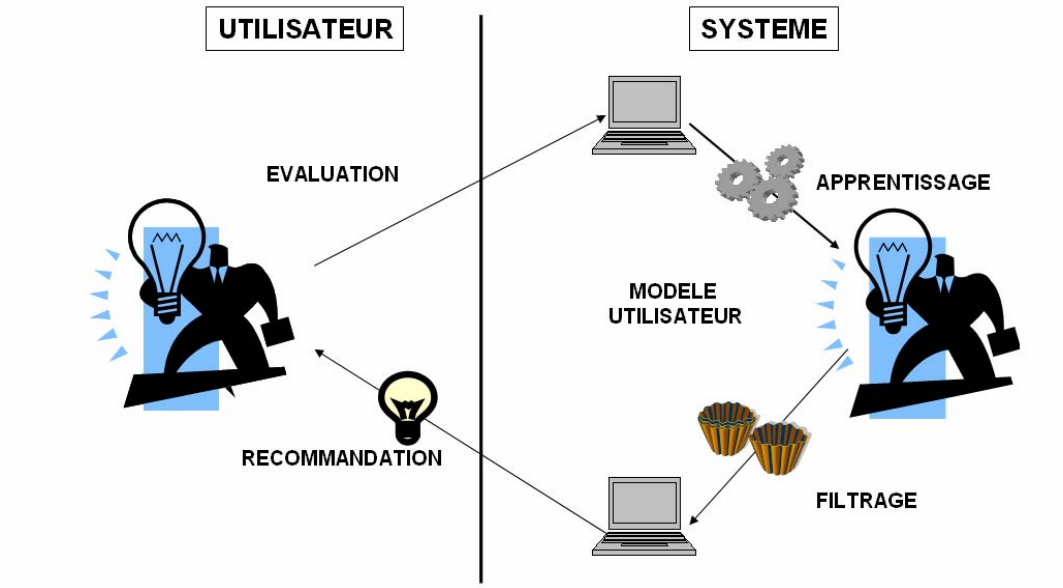

Figure 2. Modèle général pour le filtrage collaboratif d'information.

Tout d'abord, la première notion à prendre en compte est celle du taux de satisfaction des utilisateurs pour les pages qu'ils parcourent. En effet, pour caractériser un profil d'utilisateur, la première chose à comprendre, c'est si ce qu'il a lu lui a plus ou moins plu, ou pas du tout. Pour cela, deux choix sont possibles: demander à l'utilisateur de noter lui-même les pages à partir d'une échelle de notes fixée, ou bien évaluer automatiquement cette note, grâce aux informations que l'on peut récolter à partir des données du protocole de communication. En effet, chaque fois qu'un utilisateur accède à un site distant, il laisse derrière lui des traces de son passage. Pour traiter ce problème, P.K. Chan (Chan, 1999) propose une formule, pour prédire si la page proposée a été appréciée ou non. Pour cela, quatre sources générales d'informations ont été identifiées: l'historique et le « bookmark » du côté utilisateur; l' « access log » et le contenu des pages du côté serveur:

- L'historique global d'un fureteur maintient une marque du dernier moment où chaque page a été visitée. On peut donc utiliser cette donnée pour calculer combien de fois un utilisateur passe sur une page et depuis quand il n'est pas retourné la visiter.
- Chaque entrée dans un « access log » correspond à une requête HTTP, qui contient typiquement l'adresse IP du client, une marque du moment de connexion, des méthodes d'accès, une URL, un protocole, un statut, et une taille de fichier. Typiquement, une ligne de fichier de log se présente comme suit: « 216.22.34.11 - - [10/Nov/2004:04:04:54 +0200] "GET /uqam.ca/index.html HTTP/1.1" 200 2856 » :
  - ✓ 216.22.34.11 correspond à l'adresse IP du client,
  - ✓ 10/Nov/2004:04:04:54 correspond au moment de connexion,
  - ✓ GET correspond à la méthode d'accès,
  - ✓ /uqam.ca/index.html est l'URL de la page qui a été visitée,
  - ✓ HTTP/1.1 correspond au protocole de communication,
  - ✓ 200 est la valeur de retour (ici, tout s'est bien passé),
  - ✓ et 2856 correspond à la quantité d'informations transférées (généralement la taille du fichier).

  Grâce à ces informations, le temps passé sur chaque page peut être calculé. Cependant, le temps passé sur une page dépendant également de la longueur de cette page, l'intérêt d'un utilisateur pour une page sera calculé par le temps passé sur la page, normalisé par la taille de la page.

- On peut ensuite également supposer qu'une URL présente dans le « Bookmark » d'un utilisateur est considérée comme très intéressante pour celui-ci.
- Enfin, chaque page contient des liens vers d'autres pages. On suppose que si un utilisateur est intéressé par une page, il va probablement visiter les liens référencés par celle-ci. Ainsi, on conjecture qu'un fort pourcentage de liens visités à partir d'une page dénote un intérêt spécial pour cette page.

Finalement, P.K. Chan définit le degré d'intérêt d'une page par:

$$\text{Interest(Page)} = \text{Frequency(Page)} \times (1 + \text{IsBookmark(Page)} + \text{Duration(Page)} + \text{Recency(Page)} + \text{inkVisitPercent(Page)})$$

Où

- Frequency(Page), le nombre de visite de la page,
- IsBookmark(Page) = 1 si la page appartient au bookmark de l'utilisateur sinon 0
- $\text{Duration(Page)} = \dfrac{\text{TotalDuration (Page) / Size (Page)}}{\max_{\text{page} \in \text{VisitedPage}} \text{TotalDuration (Page) / Size(Page)}}$
- $\text{Recency (Page)} = \dfrac{\text{Time(LastVisit)-Time(StartLog)}}{\text{Time(Now)-Time(StartLog)}}$,
- $\text{LinkVisitPercent (Page)} = \dfrac{\text{NumberOfLinksVisited(Page)}}{\text{NumberOfLinks(Page)}}$

Les notes ainsi obtenues sont alors stockées dans la base de données utilisateur. Typiquement, on représente ces données par une matrice de notes des utilisateurs sur les pages qu'ils ont parcourues. La base de données utilisateur se présente alors sous la forme du tableau 2 suivant.

	Document 1	Document 2	Document 3	Document 4	Document 5
Utilisateur 1			7	6	
Utilisateur 2			5	6	7
Utilisateur 3			6	6	7
Utilisateur 4	7	5			7
Utilisateur 5	7	6			7

Tableau 2 : Une matrice des notes attribuées aux articles par les utilisateurs.

Formellement,

- cette matrice correspond à en un ensemble de votes $v_{i,j}$ des utilisateurs (i) sur les pages (j). Par exemple ici, on a $v_{3,3}$=6.
- v_i correspond au vecteur décrivant l'utilisateur i (l'ensemble de ses votes, $\{v_{i,j} \mid i\}$). Dans notre exemple, on a v_3=(0,0,6,6,7).
- On utilise la notation I_i pour désigner l'ensemble des pages pour lesquelles l'utilisateur i a voté. On a alors I_3={3,4,5}.
- Et nous noterons $p_{i,j}$ l'estimation du vote de l'utilisateur i sur la page j, le but étant que $p_{i,j}$ soit le plus proche possible de $v_{i,j}$.

Breese propose une classification d'algorithmes de filtrage collaboratif, exploitant ces données pour aider les utilisateurs dans leurs futures recherches (Breese et al, 1998):

- les algorithmes basés sur la mémoire utilisent l'entièreté de la base de données utilisateur pour faire des prédictions,
- alors que les algorithmes basés sur les modèles utilisent la base de données utilisateur pour estimer ou apprendre un modèle, qui sera ensuite utilisé pour les prédictions.

L'idée de base des systèmes basés sur la mémoire est que le système maintient, en premier lieu, un profil utilisateur, c'est-à-dire un enregistrement des intérêts de l'utilisateur pour certains articles. Puis, il compare ce profil avec les profils des autres utilisateurs, et pèse chaque profil en fonction de son degré de similarité avec le profil de l'utilisateur considéré. Enfin, il considère un ensemble des profils les plus similaires, et utilise l'information qu'ils contiennent pour recommander à l'utilisateur des articles qu'il n'a pas encore évalués. Pour prédire la pertinence d'un article pour un utilisateur, on calcule donc la moyenne des notes données (vote moyen pondéré, vote moyen pondéré de Bayes, etc.) aux articles par les utilisateurs ayant les mêmes goûts, en utilisant des poids différents selon la mesure de similarité entre utilisateurs (Pearson, Bayes, etc.). Par contre, dans les méthodes basées sur les modèles la tâche du filtrage collaboratif peut être vue comme le calcul de la valeur la plus probable d'un vote, étant donné ce que nous savons à propos de l'utilisateur, et le modèle que l'on aura construit à partir de la base de données utilisateur (le modèle réseau de Bayes ou le modèle cluster).

Pour conclure, ces systèmes présentent un certain nombre de limites. En effet, des problèmes subsistent pour les nouveaux documents ; ils ne peuvent être diffusés que si un minimum d'informations les concernant est collecté à partir de l'avis de l'un des utilisateurs. D'un autre côté, les

personnes ayant des goûts peu fréquents risquent de ne pas recevoir de propositions. Ces deux problèmes sont en réalité liés à la taille et à la composition de la population d'utilisateurs (le problème de la « masse critique »). En outre, ces systèmes souffrent aussi tous du problème de démarrage à froid. Les nouveaux utilisateurs commencent avec un profil vide et doivent le constituer à partir de zéro. Même avec un profil de démarrage, une période d'apprentissage est toujours nécessaire avant que le profil ne reflète concrètement les préférences de l'utilisateur. Pendant cette période, le système ne peut pas filtrer efficacement.

2.3.1.3 Le filtrage hybride

L'idée principale des approches hybrides est de permettre de tirer profit des avantages des deux approches (collaboratif et basé sur le contenu), en limitant les problèmes qui leur sont liés. En effet, nous pouvons remarquer que ces deux systèmes paraissent complémentaires. Ainsi, nous pouvons penser que combiner les deux méthodes pourrait être très bénéfique. D'où l'émergence d'approches hybrides dont l'objectif consiste à combiner les deux approches précédentes de manière efficace.
Burke (Burke, 2000) décrit sept différents types de méthodes d'hybridation: combinaison de dispositif, en cascade, utilisation de poids, etc. Cependant, toutes ces méthodes se basent sur 3 approches principales :

- Le profil pourrait alors consister en des groupes d'utilisateurs considérés comme similaires pour un thème donné (Lashkari, 1999).
- Les profils thématiques des utilisateurs sont utilisés pour les comparer dans la phase de calcul des corrélations entre utilisateurs du filtrage collaboratif (Pazzani, 1997).
- Le « boosting », qui consiste typiquement à utiliser les résultats de différentes techniques, et à les combiner (Schapire, 1998).

Les récents travaux dans ce domaine visent à développer des algorithmes hybrides cherchant à réduire le nombre de calculs «on-line» et à augmenter le nombre de calculs «off-line».

2.3.2. Architecture décentralisée

Le domaine des architectures distribuées est un sujet d'actualité. Dans cette section, nous allons nous limiter, en premier lieu, à expliquer brièvement pourquoi ce domaine suscite autant d'engouement. Ensuite, nous allons nous intéresser aux différents domaines d'application. Enfin, nous allons en expliquer la t pertinence pour notre recherche.

Il est de notoriété que l'architecture centralisée pose des problèmes de sécurité, robustesse, et de limitation de la bande passante. Les problèmes sont directement issus de l'utilisation de serveurs dont le seul but est de posséder l'annuaire des clients. Si nous voulons supprimer les serveurs centraux, il faut donc trouver le moyen de constituer un annuaire sur chaque client, puis de les faire communiquer. C'est sur ces mécanismes que sont basés les réseaux Peer-To-Peer décentralisés. Il n'y a donc plus de serveurs centraux, ce sont tous les éléments du réseau qui vont jouer ce rôle. Chaque machine dans ses rôles est identique à une autre, c'est pour cela que l'on appelle ces types de réseaux Peer-To-Peer pur. En vous connectant à de tels réseaux, vous aurez toujours besoin d'un programme mi-client et mi-serveur pour établir une connexion sur une ou plusieurs autres machines équipées, comme la vôtre, du même logiciel. Contrairement aux réseaux centralisés, où il suffisait de se connecter au serveur pour avoir accès aux informations, il faut pour avoir accès à une information :

- Apprendre la topologie du réseau sur lequel le client est connecté.
- Rechercher l'information sur tous les noeuds.
- Recevoir une réponse d'un noeud répondant aux critères.

Un grand avantage de ce nouveau type de réseaux est, en théorie, le total anonymat qu'il procure. En effet en évitant de communiquer avec une machine centralisant les demandes et les annuaires, on évite les problèmes de récupération des données utilisateur. Cependant, le principal inconvénient du Peer-To-Peer est le comportement libre de chaque Peer. En effet, dans un environnement décentralisé, chaque Peer est libre de se comporter comme il le préfère. Ainsi, des comportements de tromperie peuvent naître. Par exemple, dans les systèmes de partage de connaissances comme Kazaa, les utilisateurs peuvent fournir intentionnellement des documents erronés. Nous allons voir plus en détail ce problème dans la section 4.1.4.

Dans la revue de la littérature, nous pouvons distinguer trois classes principales d'applications Peer-To_Peer :

- Le calcul distribué ou encore « Grid Computing » : Le Peer-To-Peer peut être mis à profit pour exploiter en parallèle les capacités de calcul d'un grand nombre d'ordinateurs. En effet, ce domaine de recherche regroupe tous les systèmes capables de partager les ressources des ordinateurs à travers le réseau intranet ou Internet. Généralement, ces systèmes sont utilisés dans des applications qui nécessitent une grande puissance de calcul. De nombreux projets scientifiques exploitent cette possibilité : SETI@HOME pour la recherche de vie extraterrestre,

GENOME@HOME pour le décryptage du génome humain, etc. Ce sont des économiseurs d'écran qui utilisent la puissance de votre ordinateur et celui des autres pour analyser des données scientifiques (Foster, 2001 ; Oram, 2001).

- Le partage de connaissances : La technologie Peer-To-Peer termine d'exploiter des contenus répartis via un réseau et de mettre en oeuvre une application de gestion de connaissance. L'avantage et que la gestion de l'information se fait directement à la source, l'auteur en maîtrise directement les droits d'accès. Cependant, il en résulte la difficulté de maîtriser la qualité et la consistance du contenu. Ces applications permettent d'établir des relations d'échanges réciproques de données entre des ordinateurs distants reliés au réseau intranet ou Internet. L'exemple le plus connu est les systèmes de partage de musique et de vidéo via Internet. Les logiciels les plus connus sont Napster, Kazaa, Gnutella, E-Mule, E-Donkey, etc (Oram, 2001).
- Le travail collaboratif : C'est la principale application pour les entreprises. Le logiciel le plus connu dans ce domaine est Groove qui permet le dialogue en temps réel, le partage de documents ou d'outils dans des espaces de travail personnalisé et autonome, avec des mécanismes de synchronisation des données en temps réel et de gestion de la sécurité (authentification, confidentialité, etc.). Il existe également « Bleu » de L2T et « Magi » d'Endeavors (Frascaria, 2002 ; Merkow, 2001).

Pour conclure cette section, nous allons expliquer la pertinence d'utilisation d'une telle architecture dans notre recherche. Pour faire un rappel, une architecture décentralisée est utilisée dans des domaines d'application où la puissance de calcul réside au niveau du client et où ce dernier entre et sort du réseau constamment et sans préavis. En plus, généralement, on a recours à cette architecture dans des domaines d'application où la disponibilité du service est critique ou où les utilisateurs ne veulent pas compter sur un unique fournisseur de services. En d'autres termes, la raison principale et aussi la plus évidente d'utilisation d'un tel réseau, est d'éviter d'utiliser un serveur central. En effet, chaque Peer agit comme un serveur et un client. Ainsi, si un Peer se déconnecte du réseau, ceci diminuera, dans le pire des cas, la qualité des recommandations de documents et n'affecte pas, comme dans le cas d'une architecture centralisée où le serveur se déconnecte, le fonctionnement global du système.

2.3.3. Techniques de repérage de l'information dans les réseaux Peer-To-Peer

Récemment, un nombre important de systèmes de repérage d'information pour le réseau Peer-To-Peer ont été développés. Dans cette section, nous allons, dans un premier temps, présenter une classification des différents types de réseaux Peer-To-Peer. Ensuite, nous allons introduire brièvement les différents algorithmes qui ont été développés et publiés pour ces types de réseaux Peer-To-Peer.

2.3.3.1. Les classes de réseaux Peer-To-Peer:

En parallèle avec les réseaux Peer-To-Peer centralisés (comme Napster), il existe deux types de réseaux décentralisés (Yang et Garcia-Molina, 2001) :

- Les réseaux Peer-To-Peer non structurés : Ces réseaux, aussi qualifiés de réseaux faiblement contrôlés, se focalisent sur le partage des données et non sur les règles de leur classement. En effet, le classement des données est très faiblement contrôlé et les Peers sont incapables de décider quelles données ils veulent stocker et lesquelles ils veulent partager avec les autres Peers du réseau. Ceci réduit considérablement la disponibilité et la persistance des données.
- Les réseaux Peer-To-Peer structurés : Ces derniers, aussi appelés réseaux fortement contrôlés, sont généralement vus en tant que réseau imposant un contrôle plus strict sur le classement des données et la topologie du réseau Peer-To-Peer. Dans ce type de réseaux, les Peers sont capables de définir quelles données ils veulent stocker et lesquelles ils veulent partager. Un algorithme définit le classement des données pour chacun des Peers.

2.3.3.2. Les techniques de repérage d'information dans les réseaux Peer-To-Peer :

- Approches centralisées

Ces réseaux ne sont pas comme des réseaux Peer-To-Peer purs. En effet, ces systèmes centralisés maintiennent un index de tous les documents partagés par les Peers participants (comme Napster). Cela va sans dire que le système Peer-To-Peer centralisé utilise une base de données centrale à laquelle chaque Peer fournit un index de tous ses documents partagés. Un Peer qui recherche une

information doit automatiquement passer par la base de données centrale. En recevant la requête de recherche, le serveur recherche dans sa base de données établie avec les données envoyées par les Peers. Pour conclure, les approches centralisées sont rapides et garantissent de trouver tous les résultats possibles. Cependant, ces approches sont trop coûteuses (serveur central) et ne sont pas fiables (perte de service si le serveur tombe en panne). En plus, il y a beaucoup de liens morts dans les résultats (dépendamment de la fréquence de mise à jour de la base de données) (Yang et Garcia-Molina, 2001).

- Approches décentralisées
  - ✓ Les réseaux Peer-To-Peer non structurés :

Il existe dans ce type de réseaux deux techniques de recherche :

- Les techniques de recherche à l'« aveugle » :

Les algorithmes, qui appartiennent à cette catégorie, n'emploient aucune technique explicite pour guider la recherche.

❖ *Technique « Breadth First Search» (BFS)*

Le BFS est une technique de recherche largement répandue dans les réseaux Peer-To-Peer (Gnutella). Elle est considérée comme une approche naïve. En effet, un Peer produit un message de requête qui est propagé à tous ses voisins (Peers). Puis, quand un Peer reçoit une requête, d'abord, il fait suivre la requête à tous les Peers, autre que l'expéditeur, et ensuite, il recherche dans sa base de données locale les résultats appropriés. Enfin, si un Peer reçoit une requête et a un résultat, alors il produit un message de « QueryHit » pour transmettre le résultat. Le message de « QueryHit » inclut aussi des informations telles que le nombre de documents et de la connectivité du Peer. Comme nous pouvons le remarquer, cette technique est basée sur l'inondation du réseau entier pour la recherche d'objet. En effet, une requête est propagée le long de tous les liens et entre en contact avec tous les Peers accessibles. Bien que ce soit une technique simple à mettre en œuvre, elle sacrifie sa performance et pose un problème de surcharge du réseau (congestion du réseau). En effet, à chaque requête, il y a une consommation excessive des ressources du système. Cependant, il existe une technique, pour éviter l'inondation du réseau en entier avec des messages, qui associe à chaque requête le paramètre « Time-To-Live » (TTL). En effet, le TTL détermine le nombre maximum de sauts qu'une requête donnée devrait faire (Kalogeraki et al, 2002).

❖ *Technique «Random Breadth-First-Search » (RBFS)*

L'idée de cet algorithme est de modifier l'algorithme original du BFS de sorte que le message de requête est envoyé seulement à un nombre limité de voisins. À cet effet, un Peer fait suivre la requête seulement à une fraction de ses Peers, choisie au hasard. Le pourcentage des Peers sollicités est un paramètre de l'algorithme. L'avantage du RBFS est qu'il n'exige pas la connaissance globale ; un Peer peut prendre des décisions locales d'une façon rapide puisqu'il doit seulement choisir une partie de ses Peers. Cependant, malgré que cet algorithme réduise le nombre de messages, il reste peu fiable. En effet, avec cette méthode, des parties du réseau peuvent devenir inaccessibles (Kalogeraki et al, 2002).

❖ *Technique « Random Walkers »*

L'idée principale est que chaque Peer propage aléatoirement le message de requête à un nombre K de ses voisins. Chacun de ces Peers fait suivre la requête à n'importe lequel de ces voisins (aléatoirement). Ces messages de requêtes sont désignés sous le nom de « Walkers ». L'avantage de cette approche est qu'elle produit un nombre de messages qui est indépendant de la topologie du réseau : K*TTL. Cet algorithme ressemble à l'algorithme RBFS. Cependant, dans RBFS, l'augmentation du nombre de messages est exponentielle tandis que dans le modèle de « Random Walkers » l'augmentation du nombre de messages utilisés est linéaire. Cependant, il reste, toujours, le problème du choix aléatoire des Peers qui n'assure pas le résultat final (Lv et al, 2002).

- Les techniques de recherche dirigée :

❖ Technique « **Most Results in Past** » **(>RES).**

Dans cette technique, chaque Peer fait suivre une requête à un sous-ensemble de ses voisins choisis sur la base de certaines statistiques. En effet, cet algorithme utilise le nombre de résultats des requêtes antérieures retournés par le Peer comme heuristique pour le choix du sous-ensemble de Peers. Pour ce faire, la requête est envoyée à un nombre K de Peers qui ont retourné le plus grand nombre de résultats pour les M dernières requêtes. M et K étant des paramètres de l'algorithme (Yang et Garcia-Molina, 2002).

❖ Technique « Adaptive Probabilistic Search » (APS)

C'est une technique semblable à l'algorithme du « Random Walkers ». Dans cet algorithme, chaque Peer déploie un index local, qui calcule la probabilité relative de chaque voisin à être

choisi comme prochain saut pour une future requête. La différence principale avec l'algorithme du « Random Walkers » est que dans l'algorithme APS, le Peer utilise un feedback sur les recherches précédentes pour choisir, grâce aux probabilités, les futurs Peers, plutôt qu'un choix aléatoire. L'algorithme APS offre de meilleurs résultats que l'algorithme du « Random Walkers » (Tsoumakos et Roussopoulos, 2003).

- ✓ Les réseaux Peer-To-Peer structurés :
  - SETS (Search Enhanced by Topic segmentation:

SETS est une architecture pour la recherche efficace dans les réseaux Peer-To-Peer. Pour ce faire, le système arrange les sites participants en segments thématiques à travers le réseau. En effet, la philosophie fondamentale de SETS est d'arranger des sites dans un réseau de manière à ce qu'une requête de recherche sonde uniquement un petit sous-ensemble de sites où la plupart des documents pertinents sont localisés. Plus précisément, SETS partitionne les sites en des segments thématiques, où les documents semblables appartiennent au même segment. Chaque segment thématique a une description succincte appelée *« Topic centroid »*. Les sites sont placés dans un réseau segmenté qui se compose de deux genres de liens, des liens de longue distance relient les sites des différents segments et des liens de courte distance qui relient des sites dans le même segment. Quand une requête de recherche est lancée, elle est expédiée aux autres sites en utilisant *le protocole « topic driven routing protocol »*. Les *« Topic centroid »* sont employés pour choisir un petit ensemble approprié de segments thématiques. Après, les segments choisis sont interrogés dans l'ordre. Une requête, à un segment particulier, se propage en deux étapes : d'abord, la requête est conduite le long des liens de longue distance pour atteindre un emplacement aléatoire appartenant au segment cible. Ensuite, les liens de courte distance sont employés pour propager la requête aux sites dans le segment. Chacun des sites dans le segment font une recherchent dans leurs bases de données et les résultats de tous les sites sont combinés et sont fourni au client qui a émis la requête.

  - Keyword Relationship

Ce système est un algorithme de recherche de documents par mots clés dans un réseau Peer-To-Peer qui améliore le repérage de l'information désirée. L'idée principale est une expansion de la requête. En d'autres termes, le système ajoute des mots-clés appropriés à la requête originale. Cette expansion de la requête est basée sur une base de données de relations entre mot-clés (KRDB), qui est contrôlée dans un mode distribué par les Peers participants. Le KRDB est amélioré par la recherche et des procédés de repérage de l'information (Morikawa et al, 2003).

- Swarm Search

Cet algorithme de recherche est basé sur l'imitation du comportement de colonies de fourmis. Il emploie un nombre élevé d'agents avec des comportements simples, et génère une espèce d'agents (un type de l'agent) pour chaque requête utilisateur. Les agents de la même espèce ont l'unique mission de rechercher des chemins de ressource se conformant aux critères donnés par le profil d'une demande d'utilisateur spécifique. Cet algorithme fournit une méthode générale pour produire des solutions aux problèmes combinatoires d'optimisation multicritères (Wittner, 2002).

2.3.4. Système multi-agent

Dans cette section, nous allons, dans un premier temps, présenter brièvement le lien naturel qui existe entre les systèmes Peer-To-Peer et les systèmes multi-agent. Ensuite, nous allons justifier l'intérêt de la présence des comportements sociaux complexes dans les systèmes multi-agent.

Nous allons considérer le système de filtrage comme un ensemble d'agents indépendants, un agent pour chaque utilisateur du système. Ainsi, le système est constitué par une population d'agents autonomes en interaction où chaque agent possède les capacités de communication, de coordination et de collaboration. Cependant, nous aurions besoin d'un réseau pour établir la communication entre les agents. Pour ce faire, nous allons utiliser l'approche Peer-To-Peer pour relier les agents entre eux. En effet, de récentes études ont été conduites pour essayer de combiner la technologie Peer-To-Peer et les systèmes multi-agents. Les chercheurs, dans ce domaine, voient le Peer-To-Peer comme paradigme pour les réseaux de systèmes autonomes qui peuvent se joindre (se connecter) au réseau ou en partir (se déconnecter) à tout moment. En effet, nous avons pu voir que, dans un réseau Peer-To-Peer, chaque Peer possède une autonomie significative, et qu'un Peer peut se joindre à différents groupes de Peers afin de réaliser différents buts. En plus, certains concepts du Peer-To-Peer, tel que le groupe de Peer et l'autonomie, sont des concepts typiques des systèmes multi-agents. Enfin, le concept de groupe de Peers est bien convenable pour soutenir l'aspect du dynamisme des organisations sociales au sein des systèmes multi-agents. Donc, il est approprié, dans notre travail, de proposer fondamentalement le Peer-To-Peer comme manière de déployer un système multi-agents ouvert.

Dans un autre ordre d'idée, l'un des problèmes cruciaux des SMA est celui de l'organisation des agents. Au mépris de l'importance de cette dernière, elle reste jusqu'à présent une notion mal définie. L'organisation au sein d'un SMA implique que ces derniers réussissent à se coordonner, collaborer,

gérer l'accès aux ressources, résoudre les conflits, échanger des informations en fonction du but à atteindre. Pour ce faire, des contraintes sont incorporées aux comportements individuels des agents. Ainsi, nous pouvons limiter leur marge d'action dans le but d'obtenir une certaine forme de coordination, donnant lieu à un comportement collectif intéressant. Comme le suggère Almgreen, la mise en situation de ces agents dans un environnement social plus dense donnerait peut-être un avantage plus grand à la communication. L'auteur propose un environnement plus hostile comme exemple (Almgreen, 2000). En outre, selon Mataric, le comportement social maximise le bénéfice individuel moyen en maximisant le bénéfice collectif. Dans ces recherches, l'auteur montre en particulier la pertinence et l'efficacité des règles sociales particulièrement altruistes (Mataric, 1994 Mataric, 1997).

2.4. Proposition de recherche

Compte tenu de ce qui précède, nous nous sommes proposé d'approfondir quelques points de recherche. En premier lieu, nous allons nous intéresser aux systèmes de filtrage l'information hybride. En effet, ce dernier, collaboratif et individuel en même temps, donc évolutif, est un domaine de recherche en plein essor. En outre, nous allons nous pencher sur l'incorporation d'une architecture Peer-To-Peer et les solutions au problème, introduit brièvement dans la section 2.3.2, de tromperie. Enfin, nous allons explorer l'adoption de comportements sociaux complexes dans des systèmes multi-agent.

Pour ce faire, le modèle cognitif que nous proposons est une plate-forme d'agents collaboratifs pour le filtrage hybride de l'information basée sur une architecture Peer-To-Peer. En effet, nous allons considérer le système comme un ensemble d'agents indépendants, un agent pour chaque utilisateur du système. Il s'agit d'un système où l'utilisateur est passif et où les informations lui sont délivrées automatiquement (la technologie « Push » : comme les services d'alerte).

2.4.1. Motivations spécifiques de recherche

Relativement à notre proposition, la motivation principale de notre recherche est que la collaboration peut améliorer la performance du système et la qualité des résultats obtenus.

2.4.2. Questions spécifiques de recherche

Compte tenu de notre proposition, nous allons nous interroger sur (1) les méthodes pour combiner le filtrage basé sur le contenu et le filtrage collaboratif afin d'améliorer la performance et la qualité, et de résoudre les problèmes (comme exemple, le problème de la masse critique) ; (2) comment l'absence de serveur central, pour la gestion des profils d'utilisateurs, affectera-t-elle le système de filtrage ; (3) le problème de tromperie dans la recommandation des utilisateurs ; (4) l'intégration de comportements sociaux complexes dans les fonctionnalités de recommandation.

2.4.3. Objectifs spécifiques de recherche

L'objectif principal de ce travail est de soutenir l'utilisateur dans son processus d'accès à l'information. Ce dernier peut être divisé en quatre sous-objectifs :

- ✓ Le développent d'un algorithme pour combiné filtrage basé sur contenu et le filtrage collaboratif.
- ✓ L'étude des comportements d'un utilisateur face à un document pour être employée comme paramètre dans le système de filtrage (comme exemple l'évaluation des documents).
- ✓ La création d'une plate-forme de filtrage qui emploie une architecture décentralisée.
- ✓ L'emploient d'un système de recommandation pour recommander des personnes plutôt que des documents et ainsi, transformer le système en un repère d'experts et augmenter la confiance au sein du système de recommandation.

3. Volet cognitif du projet de recherche.

3.1. Les phénomènes sociaux

Le but du volet cognitif de notre projet de recherche est, en premier lieu, d'identifier quelques comportements sociaux chez les primates susceptibles d'être reproduits dans un système multi-agent. Ensuite, en suivant l'axe éthologique qui consiste à étudier des sociétés naturelles complexes, en particulier les sociétés de primates, nous proposons de construire des modèles de cognition sociale capables de reproduire, dans un système multi-agents, des phénomènes sociaux naturels. En d'autres termes, nous allons mettre en œuvre ces phénomènes sociaux dans les interactions entre agents au sein de notre système.

Le domaine dans lequel s'inscrit l'étude des comportements sociaux au sein des sociétés naturelles complexes est celui de l'éthologie. Cette dernière étudie le comportement animal et humain. Plus précisément, l'éthologie s'intéresse à l'ensemble des facteurs qui vont faire que tel animal va exprimer tel comportement. Ainsi, une revue de la littérature dans ce domaine nous a permis d'identifier quelques phénomènes sociaux naturels capables de nous intéresser :

- La représentation des relations existant entre les congénères. Ainsi, chez les primates, les individus d'un groupe se connaissent mutuellement ainsi que les relations sociales existantes (dominance, parenté, affiliation). En plus, les individus peuvent tenir compte dans leurs comportements de la transitivité de la hiérarchie (si A domine B et B domine C, alors généralement A domine C). Ceci implique que les primates disposent d'une représentation symbolique des relations sociales, et d'un mécanisme d'inférence (Cheney et Seyfarth, 1980 ; Dasser, 1988).
- Des comportements d'agrégation : les comportements les plus courants chez les primates sont les comportements « affiliatifs » (épouillage mutuel, jeu, etc.). L'objectif essentiel de ces derniers est le renforcement des liens sociaux. En plus, les individus sont portés à demander leurs anciens partenaires (stratégie de type « donnant donnant ») (De Waal, 1982).
- Des comportements de coopération : ces derniers consistent dans leur capacité à former des coalitions pendant les combats, et à utiliser des réseaux d'alliance pour le maintien (macaques, babouins) ou le renversement (chimpanzés) de l'ordre établi. Ces réseaux se constituent et évoluent en fonction des liens de parenté, mais aussi des interactions affiliatives entre individus (Drogoul et Collinot, 1997).
- Des comportements de coordination : les techniques de chasse en groupe chez les chimpanzés, décrites par (Boesch, 1989), en sont l'illustration parfaite. En effet, les chasseurs encerclent leurs proies et anticipent leurs déplacements. Pour ce faire, la coordination doit être gérée de manière parfaite entre les participants.
- L'allocation dynamique de tâches en fonction des besoins du groupe et des compétences individuelles : la chasse en fournit également un exemple. En effet, les chimpanzés doivent jouer un certain nombre de rôles (rabatteur, poursuivant, etc.) bien définis, qu'ils doivent se répartir dynamiquement en fonction du nombre de chasseurs, de l'évolution de la chasse, et des compétences de chacun (Drogoul et Collinot, 1997).

- La réduction des tensions sociales : (De Waal, 1982) explique comment les tensions sociales engendrées par les conflits et la hiérarchie sont éliminées. Dans un SMA capable d'utiliser dans son fonctionnement même des conflits entre agents (Galliers, 1990), il est nécessaire d'éviter que les tensions ne s'accumulent de manière excessive, sous peine de devenir le problème central des agents, au détriment de leur tâche collective initiale.

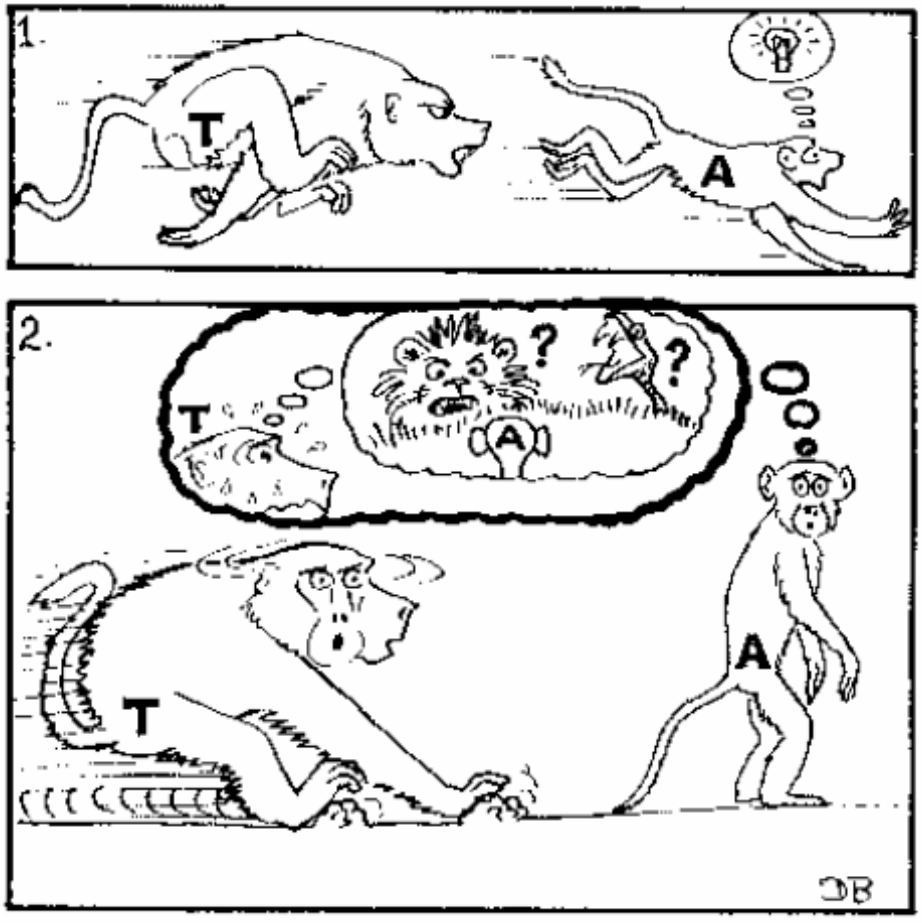

Figure 3 : Un exemple de tromperie sociale
d'après (Byrne, 1995)

- La « tromperie sociale » : observée chez certaines espèces de primates, elle consiste à manipuler les congénères à travers l'utilisation « hors contexte » de comportements habituels (Byrne et Whiten, 1988). Il s'agit donc en fait d'une manipulation liée à la « sémantique » des comportements, ou plus précisément à une connaissance de l'effet que peuvent avoir les comportements d'un agent sur ses congénères. La figure 3 présente un exemple de tromperie

sociale. En effet, un mâle A est poursuivi par un adulte T. Mais au lieu de continuer à courir, A se dresse soudain sur ses pattes postérieures et fixe l'horizon, comme le fait tout mâle qui a perçu un danger. Son poursuivant s'arrête et regarde dans la même direction. Malgré qu'il n'y ait rien de suspect, la poursuite s'arrête (Byrne, 1995).

3.2. L'émergence d'organisation

La notion d'émergence est présente, dans la littérature, depuis longtemps. Marvin Minsky, définit ce concept comme suit : « Apparition inattendue, à partir d'un système complexe, d'un phénomène qui n'avait pas semblé inhérent aux différentes parties de ce système. Ces phénomènes émergents ou collectifs montrent qu'un tout peut être supérieur à la somme de ses parties » (Minsky, 1986). En d'autres termes, on peut définir l'émergence comme l'idée qu'il existe dans un système des propriétés présentes à un certain niveau d'organisation qui ne peuvent être déduites des propriétés de niveaux inférieurs (voir figure 4).

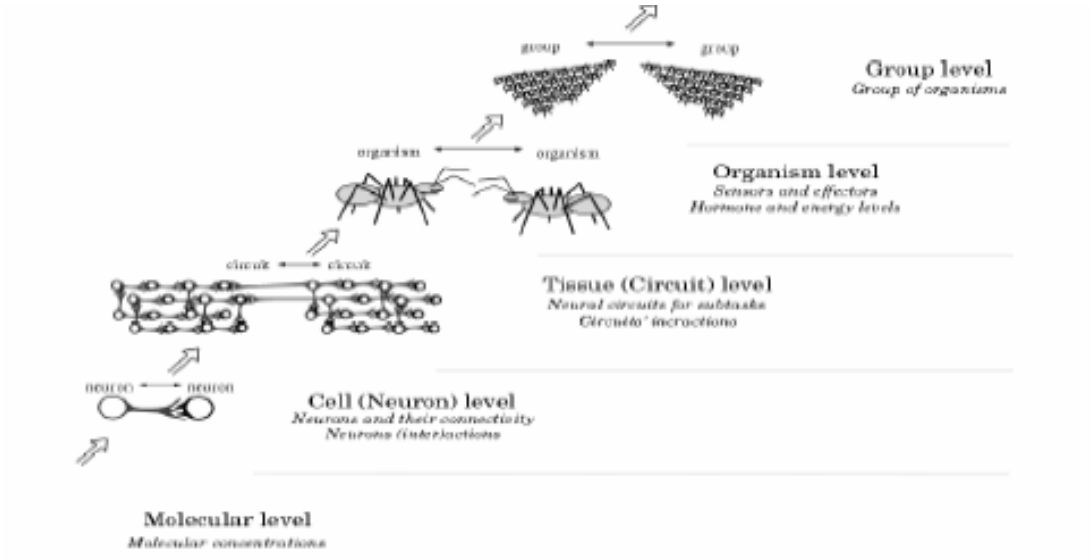

Figure 4 : Les niveaux hiérarchiques des organismes
d'après (Vaario, 1994).

Dans le domaine des systèmes multi-agents, la définition la plus utilisée est la généralisation du calcul émergent proposée par Müller (Müllller, 1998). Un phénomène est dit émergent si:

- il y a un ensemble d'agents en interaction entre eux et via l'environnement dont la dynamique n'est pas exprimée dans les termes du phénomène émergent à produire, mais dans un vocabulaire ou une théorie D;
- la dynamique des agents en interaction produit un phénomène global qui peut être une structure stable, une trace d'exécution ou n'importe quel invariant statique ou dynamique ;
- ce phénomène global est observé soit par un observateur extérieur, soit par les agents eux-mêmes en des termes distincts de la dynamique sous-jacente, c'est-à-dire avec un autre vocabulaire ou théorie D'.

Dans le reste de cette section, nous allons nous intéresser à l'émergence d'organisation résultante des interactions accomplies entre agents. En effet, dans le domaine des systèmes multi-agents, de nombreux travaux ont montré, par des expériences empiriques, que des organisations pouvaient émerger des interactions entre agents. Par exemple, dans la simulation de robots explorateurs (Drogoul et Ferber, 1992b), les phénomènes de pistes ou de chaînes émergent des coordinations locales qui ont lieu entre les individus. Dans le même ordre d'idées, dans la simulation des comportements au sein d'une fourmilière (Drogoul et Ferber, 1994b), les interactions individuelles donnent naissance à des phénomènes d'organisation du travail et de structuration des rôles. Enfin, dans les expériences « talking heads » (Steels, 2003a ; Steels, 2003b ; Steels, 2000), les interactions individuelles permettent la formation automatique et l'apprentissage individuel et en groupe du langage. Cependant, nous allons nous concentrer sur les phénomènes qui émergent des interactions individuelles au niveau des trois aspects de coopération suivants :

- Le regroupement : il consiste à rapprocher les agents, soit physiquement, soit via un réseau de communication.
- La spécialisation : elle associe aux agents des rôles temporaires par une sorte d'adaptation individuelle. Cette spécialisation peut être bénéfique à la collectivité en augmentant la capacité du groupe à résoudre plus rapidement un problème.
- Le partage des tâches et des ressources : il permet à des agents de se répartir des tâches, des informations et des ressources de manière à réaliser un objectif commun.

4. Mise en œuvre de la proposition

Dans cette étape, les détails de mise en oeuvrent de notre modèle. Pour ce faire, nous allons présenter en détail la méthode choisie pour combiner le filtrage basé sur le contenu et le filtrage collaboratif. Nous allons aussi voir comment le choix d'une architecture décentralisée affectera la structure du système de filtrage.

4.1. Filtrage collaboratif via le contenu

Lors de notre revue de la littérature, nous avons constaté que prendre en compte la description des utilisateurs améliore les résultats (Filtrage collaboratif). De la même façon, nous avons constaté que prendre en compte la description des articles améliore les résultats (Filtrage basé sur le contenu). Ainsi, nous envisageons donc d'utiliser le contenu des documents lus par les utilisateurs pour aider dans le filtrage collaboratif.

Il faut nous rappeler que les deux approches donnent des résultats très différents selon le but de l'utilisation du profil utilisateur. Le filtrage collaboratif peut s'avérer très utile pour conseiller les utilisateurs sur certains textes qu'ils n'ont pas lus, et dont on sait qu'ils intéressent son groupe. En revanche, déterminer un profil thématique pour chaque utilisateur permet par contre d'évaluer si un texte est pertinent ou non pour cet utilisateur, en fonction de ses centres d'intérêt identifiés. Or, ces deux aspects du profil paraissent importants. Connaître les goûts thématiques personnels paraît la meilleure méthode pour fournir une aide spécialisée au client, mais associer les clients entre eux permet de faire bénéficier chacune des opinions des autres.

Le filtrage collaboratif tel que nous l'avons présenté précédemment utilise l'ensemble des votes des utilisateurs sur les articles pour leur attribuer un degré de similarité. La matrice de données utilisée ressemble alors au tableau 3 suivant:

	Article 1	Article 2	Article 3	Article 4	Article 5 considéré
Utilisateur 1			7	6	???
Utilisateur 2			5	6	1
Utilisateur 3			6	6	3
Utilisateur 4	7	5			6
Utilisateur 5	7	6			4

Table 3 : Filtrage collaboratif basé sur les notes

Ici, nous cherchons la note que l'utilisateur 1 va attribuer à l'article 5, en fonction des notes que les autres utilisateurs ont attribuées à l'article, et de la similarité entre leurs vecteurs de notes sur les autres articles.

L'idée du filtrage collaboratif via le contenu est d'utiliser les profils thématiques des clients, formés à partir de l'analyse du contenu des articles qu'ils ont lus, pour comparer les utilisateurs dans la phase de calcul des corrélations entre utilisateurs du filtrage collaboratif (Pazzani, 1997). La matrice de données utilisée ressemble alors au tableau 4 suivant:

	Mot 1	Mot 2	Mot 3	Mot 4	Mot 5	Article 5 considéré
Utilisateur 1	2.5	0	0.2	0	0	???
Utilisateur 2	1.1	0	1.1	1.5	0	1
Utilisateur 3	1.5	0	3.5	1.5	0.5	3
Utilisateur 4	1.1	1.1	2.1	2.0	2.5	6
Utilisateur 5	1.1	2.2	0	0	3.5	4

Table 4 : Filtrage collaboratif basé sur le contenu

Cette fois, nous cherchons la note que l'utilisateur 1 va attribuer à l'article 5, en fonction des notes que les autres utilisateurs ont attribuées à l'article, et de la similarité entre leurs vecteurs de mots, représentant l'ensemble des thèmes identifiés comme préférés (récurrents) dans leurs lectures. Il nous reste, maintenant, qu'à identifier les paramètres de notre approche.

4.1.1. Méthode d'évaluation des documents

L'évaluation se fera sans la participation de l'utilisateur. Nous voulons éviter à l'utilisateur de s'investir dans la définition de son profil, au lieu de lui demander si une page lui a plu ou non, on peut le deviner, en se servant des informations que l'on peut obtenir lors de son passage sur le site. En effet, chaque fois qu'un internaute visite une page d'un site web, il laisse derrière lui des traces de son passage. Pour ce faire, nous allons utiliser le degré d'intérêt d'une page défini par Chan et introduit dans la section 2.3.1.3 (Chan, 1999):

$$Interest(Page) = Frequency(Page) \times$$
$$(1 + IsBookmark(Page) + Duration(Page) + Recency(Page) + LinkVisitPercent(Page)),$$

4.1.2. Niveau d'analyse appliqué au texte

Nous allons opter pour une approche en profondeur qui indexera, après une phase de prétraitement du texte, l'intégralité du texte. C'est approche dite « bag-of-words » où nous considérons le texte comme un sac de mots et en supposant que tous les mots indépendants les uns des autres.

4.1.3. Représentation adoptée pour le profil thématique

La représentation adoptée pour le profil thématique est composée d'un profil initial rempli par l'utilisateur lui-même (pour régler le problème de la masse critique); puis, d'un profil thématique à long terme, représenté sous la forme de vecteurs des mots les plus fréquents dans l'ensemble des lectures du client; et enfin d'un profil thématique à court terme, représenté sous la forme de vecteurs des mots les plus fréquents dans les dix dernières lectures du client. Avec cette représentation, l'utilisateur ne serait donc pas enfermé dans un profil figé, mais l'ensemble des thèmes abordés serait toujours conservé en mémoire.

4.1.4. Réseau de compétence

Comme nous l'avons dit plus haut, notre étude s'inscrit dans le cadre d'un système multi-agent. D'ailleurs, Russel et Norvig définissent les agents comme suit: « An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors. » (Russell et Norvig 1995). Donc, un agent est situé dans un environnement qu'il peut percevoir et sur lequel il peut agir. Pour ce faire, dans le cadre de notre projet, nous proposons que chaque Peer possède une vue, soit directe, soit indirecte, sur les autres Peers. Cette vue reflète l'importance sociale du Peer et évolue en fonction des interactions observées. Nous pensons prendre en compte la performance, la confiance, et la proximité comme paramètre pour cette vue. Dans le but d'étudier les relations sociales émergentes à partir de ces vues au niveau du regroupement et de la spécialisation des Peers, et au niveau du partage des ressources.

Pour ce faire, nous nous proposons d'introduire une notion importante, à savoir la compétence. Cette dernière nous permettra de gérer les compétences de chaque agent et ainsi, faire émerger son statut d'expert ou pas. Plutôt qu'ignorer la compétence des agents, autant la mettre en valeur, et ainsi dessiner une cartographie relationnelle du réseau de compétences. Le but est de transformer les systèmes en un repère d'experts et de résoudre le problème de tromperie. Pour ce faire, il faut recommander des personnes au lieu des documents.

La structure du réseau de compétences est composée de poids de compétence représentant l'expertise d'un nœud selon un autre. Par exemple, dans la figure 5, l'utilisateur 2 possède une expertise de 0,2 selon l'utilisateur 1. Ainsi, chaque utilisateur possède un vecteur contenant un ensemble de poids de compétence évaluant la compétence des autres utilisateurs.

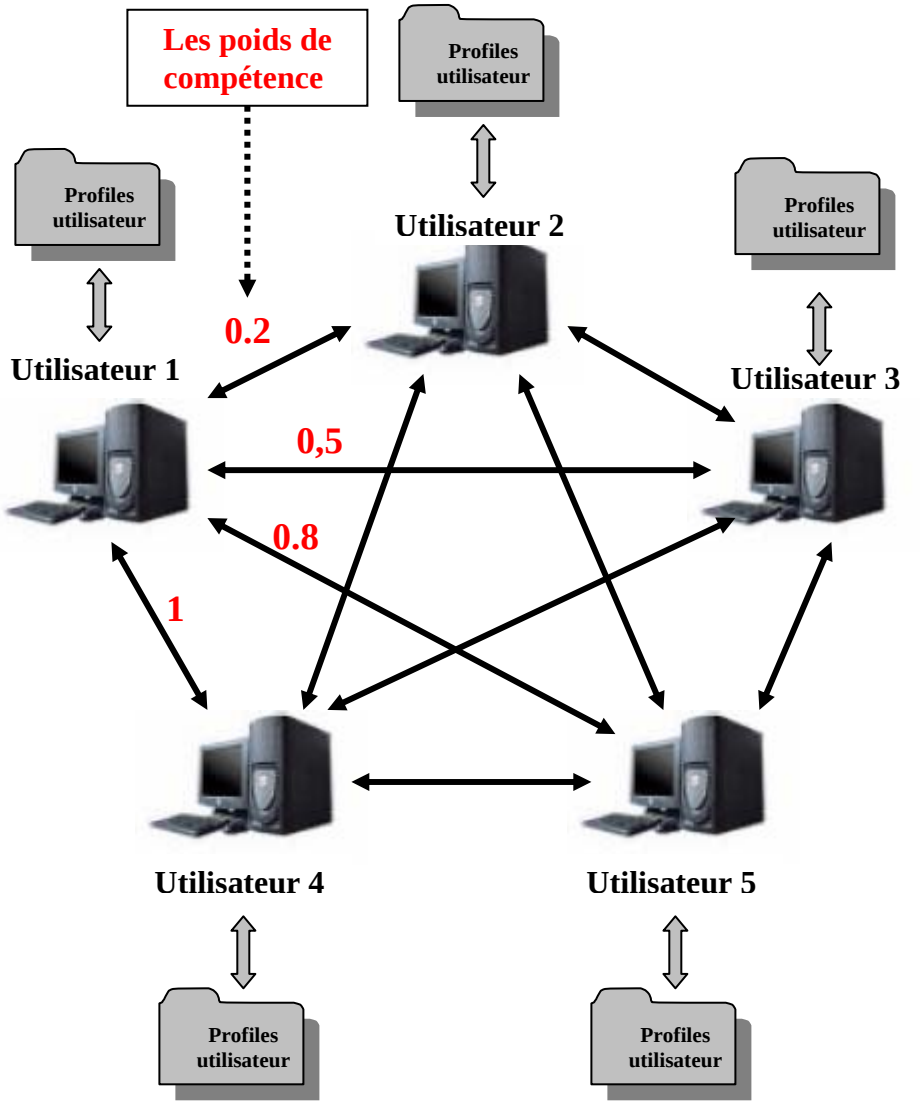

Figure 5 : La structure du réseau de compétences.

Les poids de compétence sont calculés en fonction de trois nouvelles variables dont les valeurs varieront entre [0, 1] :

- La proximité : Elle représente la distance qui sépare le nœud client du noeud serveur. Pour la calculer, il faut apprendre la topologie du réseau sur lequel le client est connecté (voir figure 6).

  - ✓ Le client A se connecte sur le réseau, mais il ne connaît pas la topologie du réseau (A est totalement aveugle).
  - ✓ Pour connaître les autres membres du réseau, A va émettre une demande d'identification des noeuds du réseau.
  - ✓ Les noeuds recevant la demande vont à leur tour la répercuter sur tous les noeuds voisins et ainsi de suite (comme les noeuds B, C et D).
  - ✓ Lorsque que la trame est reçue et identifiée par un autre client, le noeud renvoi une trame d'identification à A.
  - ✓ Ainsi A va peu à peu pouvoir identifier tous les noeuds du réseau et se créer un annuaire.

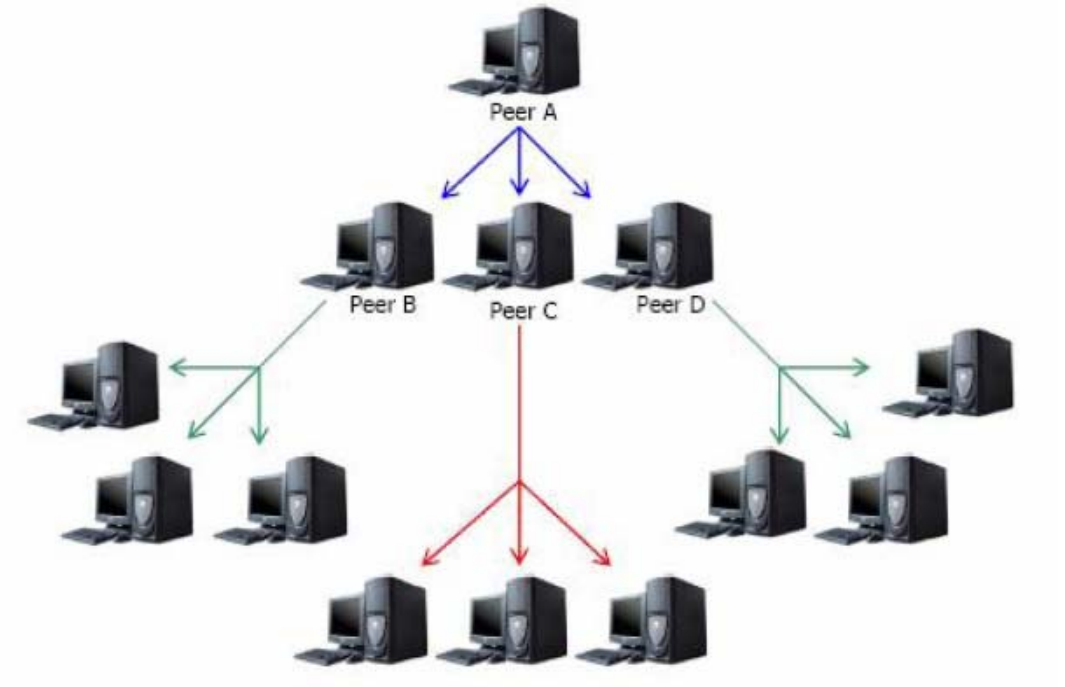

Figure 6 : La topologie du réseau.

- La performance : Cette variable est très simple mais très importante pour la performance de l'intégralité du système. En effet, la performance d'un noeud sera calculée en fonction du temps de réponse.
- La confiance : La confiance n'est pas un concept artificiel mais un concept très humain et social. L'homme a étudié le concept de confiance depuis les premiers philosophes. Il existe de nombreuse définition de la confiance. Cependant, la définition de Gambetta est la plus utilisée : « ... trust (or, symmetrically, distrust) is a particular level of the subjective probability with which an agent assesses that another agent or group of agents will perform a particular action, both before he can monitor such action (or independently of his capacity ever to be able to monitor it) and in a context in which it affects his own action. ».

  Alfarez et Hailes identifient trois genres de confiance dans le contexte des communautés virtuelles (Alfarez et Hailes, 2000). La confiance *interpersonnelle* est la confiance directe qu'un participant a dans d'autres. La confiance *impersonnelle* représente la façon dont le participant perçoit le système auquel il participe. La confiance de *disposition* est l'attitude de confiance générale du participant.

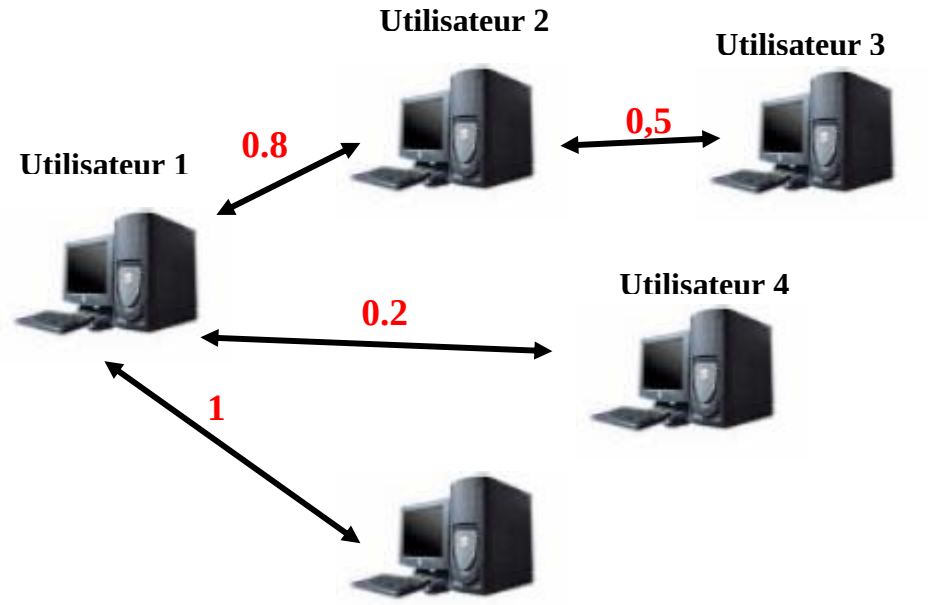

Figure 7 : Transitivité des poids de confiance.

L'utilisateur décide des valeurs de confiance d'un nœud. Cependant, la manière de décision d'un utilisateur des valeurs de confiance peut dépendre de beaucoup de facteurs. En plus, dans le cas d'un nouvel utilisateur, une première estimation peut être calculée par transitivité. Si par exemple, dans la figure 7 l'utilisateur 1 recherche le poids de confiance de l'utilisateur 3, alors le poids de confiance de l'utilisateur 1 sur l'utilisateur 3 sera égal au produit du poids de confiance de l'utilisateur 1 sur l'utilisateur 2 par le poids de confiance de l'utilisateur 2 sur l'utilisateur 3, à savoir 0,8 *0,5 qui nous donne un poids de confiance de 0,4.

Pour conclure, tout comme les interactions entre êtres vivants, l'expert sera celui qui a le poids de compétence le plus fort. Si celui-ci perd sa crédibilité (variable de confiance), il perdra son statut d'expert, n'étant plus celui qui a le poids le plus fort, tout comme dans la vie réelle. Ainsi, il y aurait une émergence des relations sociales à partir des comportements observés lors des interactions entre individus.

4.2. Architecture du système.

Beaucoup de recherches ont été consacrées aux solutions pour les systèmes centralisés de recommandation. Les systèmes traditionnels de recommandation possèdent de grands serveurs. Les utilisateurs ne peuvent interagir qu'à travers le serveur.

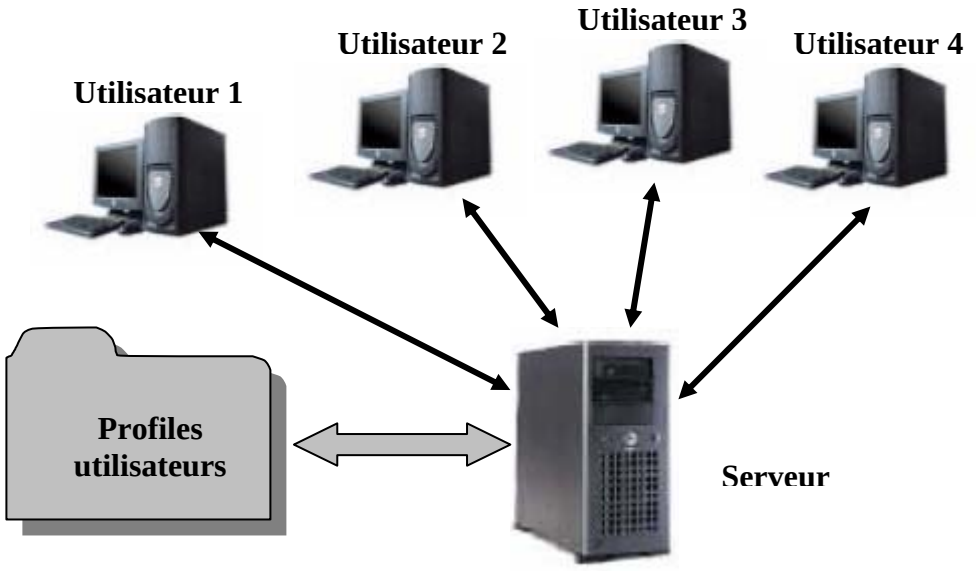

Figure 8 : Architecture générale de systèmes centralisés de recommandation.

Le serveur rassemble l'information des préférences des utilisateurs sur des documents par leur interaction, implicite ou explicite, avec le serveur. Basé sur des ces profils, le serveur donne des recommandations concernant quels articles l'utilisateur aimera. La figure 8 nous montre l'architecture générale de systèmes centralisés de recommandation.

Il est évident que cette architecture présente plusieurs inconvénients. En effet, un système de recommandation doit stocker beaucoup de données de beaucoup d'utilisateurs. En outre, si le serveur du système de recommandation est en panne ou non accessible, l'utilisateur ne peut pas utiliser le service. En plus, les politiques de sécurité et de protection des renseignements privés sont définis par le serveur du système de recommandation. Pour répondre à tous ces problèmes, nous avons opté pour une approche décentralisée.

Cependant, Il y a deux propriétés qui affectent l'architecture d'un système décentralisé de recommandation : là où l'information est stockée et où les recommandations sont calculées. Pour ce qui nous concerne nous avons opté pour système de recommandation purement décentralisé. En effet, l'information est stockée et les recommandations sont calculées au niveau des clients. Cependant, nous pouvons imaginer, dans de futurs travaux, que si un agent ne possède pas assez de ressources (mémoires, CPU), il peut demander des recommandations à un autre agent dans lequel il a confiance. Dans ce cas, il y aurait une émergence de Peers experts (socialement plus importants). Ces derniers se spécialisent comme des serveurs pour des clients de niveau inférieur. En d'autres termes, il y aurait une émergence de relations hiérarchiques entre les Peers. La figure 9 nous montre l'architecture générale de systèmes décentralisés de recommandation.

5. Démarche méthodologique adoptée.

Comme nous l'avons mentionné plus haut, ce travail est une étude exploratoire ayant pour but de développer un modèle multi-agent pour le filtrage collaboratif de l'information. Selon Sekaran (Sekaran, 1992), une étude exploratoire est entreprise lorsque nous ne connaissons pas la problématique courante ou que nous n'avons pas d'information sur la façon de résoudre des problèmes similaires.

Pour pallier aux problèmes reliés à la recherche exploratoire, nous avons opté pour l'utilisation d'une adaptation du cadre de Basili (Basili, Selby et Hutchens, 1986). Ce dernier fournit une méthode pour classifier d'une manière systématique des étapes du projet aidant ainsi à mieux comprendre le travail et faciliter l'évaluation des études expérimentales. Pour ce faire, ce modèle comporte 4 phases principales : la définition, la planification, l'exécution et l'interprétation des résultats.

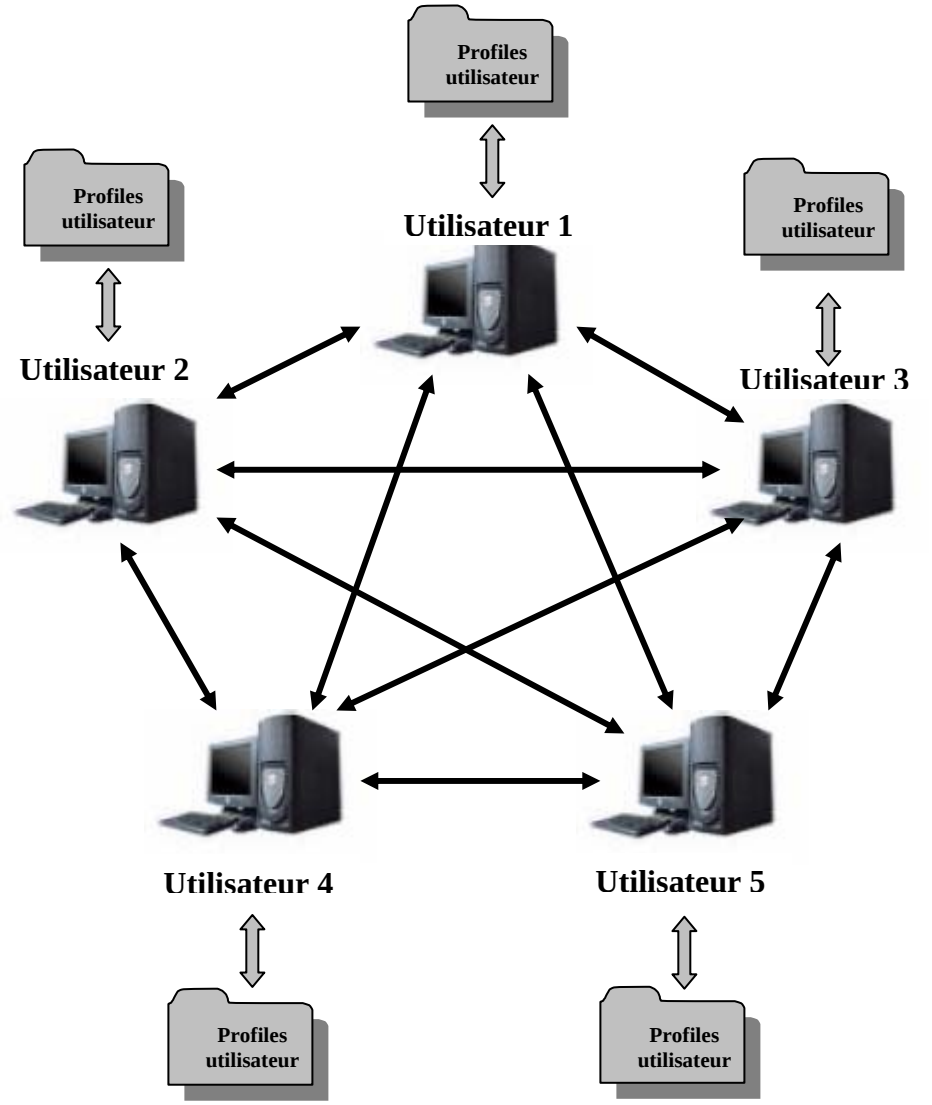

Figure 9 : Architecture générale de systèmes décentralisés de recommandation.

Dans ce travail, nous avons utilisé l'adaptation du cadre de Basili proposée par Abran et son équipe (Abran, Laframboise et Bourque, 1999). La structure du cadre de Basili, reposant sur les 4 phases principales demeures inchangées (Annexe 1). Cependant, quelques adaptations sont réalisées à l'intérieur de chaque phase. La phase de définition comporte 5 composantes (Voir la section 2.2) :

- La motivation : cette composante présente la raison qui pousse l'auteur à entreprendre le projet ; Elle identifie la grande question à laquelle on veut répondre.
- L'objet : il détermine l'entité principale à être étudiée.
- L'objectif : il est le but précis du projet ; il présente le véritable problème qu'on veut résoudre.
- Les utilisateurs : ils sont les personnes susceptibles de mettre en pratique les résultats de la recherche.

En se basant sur une revue approfondie de la littérature, la phase de planification identifie les étapes du projet qui seront réalisées dans la phase d'exécution. L'interprétation des résultats du projet constitue la dernière phase du projet. Elle est subdivisée en trois parties. En premier lieu, le contexte d'interprétation dans lequel, nous examinerons le résultat final de la recherche pour voir si l'objectif a été atteint ou non. Ensuite, l'extrapolation dans laquelle, nous étudierons les résultats pour voir s'ils peuvent être utilisés dans d'autres circonstances. Enfin, dans les travaux subséquents, nous discuterons des améliorations possibles.

Pour résumer, le cheminement méthodologique, basé sur le cadre de Basili, proposé sera organisé comme suit :

phases principales du cadre de Basili	Le cheminement méthodologique
La définition	Identification de la problématique Etude de l'existant
La planification	Proposition d'un modèle
L'exécution	Implémentation du prototype
L'interprétation des résultats	Evaluation du prototype

Tableau 5 : le cheminement méthodologique, basé sur le cadre de Basili.

6. Évaluation du projet de recherche.

L'objectif de l'évaluation de notre projet de recherche est d'identifier, en premier lieu, l'impact du système selon trois points de vue :

- notre modèle amène-t-il une amélioration de la performance et de la qualité des documents résultants de la méthode de filtrage-repérage de l'information?
- est-ce que l'utilisateur tire profit du système ?
- est-ce qu'il y a un recouvrement plus complet de l'espace d'information ?

- est-ce que notre système offre un bon rapport qualité/performance ? En d'autres termes, est-ce que nous avons pu trouver un bon compromis entre la qualité des documents et la performance du système (temps réponse, consommation de ressources système, etc.).

Ensuite, de dégager des améliorations possibles selon les résultats de notre évaluation.

Pour ce faire, l'évaluation du modèle du point de vue du filtrage de l'information sera effectuée selon les règles de l'art du domaine. Nous envisageons réaliser des expérimentations à partir des collections de la conférence TREC (Text REtrieval Conference ; http://trec.nist.gov/). Ceci permet d'effectuer des comparaisons systématiques avec les résultats déjà publiés dans les conférences passées sur les mêmes collections pour les mêmes tâches. Une adaptation des méthodes traditionnelles doit être étudiée afin de prendre en compte le contexte collaboratif de la tâche. À cet effet, nous allons opter pour deux types d'expérimentations. Dans un premier temps, nous allons procéder à des expérimentations « off-line » (des simulations) en utilisant un groupe d'utilisateurs virtuels et en générant les paramètres aléatoirement (évaluation, confiance, performance, etc.). Dans un deuxième temps, nous allons procéder à des expérimentations «on-line» (réel) avec un groupe d'utilisateurs.

Les résultats de nos expérimentations devront dégager deux types de variables. D'une part, nous allons mesurer la performance. Pour ce faire, dans un premier temps, nous proposons utiliser le temps réponse comme variable pour la performance. D'autre part, la qualité des documents doit être mesurée. Étant donné que la qualité d'un document correspond au degré d'intérêt de l'utilisateur (introduit dans la section 2.3.1.3) à ce dernier (Chan, 1999), nous allons utiliser ce dernier pour calculer la qualité des documents.

$$Interest(Page) = Frequency(Page) \times (1 + IsBookmark(Page) + Duration(Page) + Recency(Page) + LinkVisitPercent(Page)).$$

7. État d'avancement du projet de recherche.

Jusqu'à présent, nous avons pu identifier la problématique principale de notre projet de recherche et nous avons pu procéder à une revue de la littérature qui nous a permis d'identifier les pistes de recherche les plus pertinentes pour répondre à celle-ci. Nous avons également pu formuler une ébauche d'un modèle cognitif et d'une architecture de la solution proposée. A cet effet, nous avons pu définir nos protocoles, nos fonctionnalités, nos algorithmes et notre structure.

Pour pouvoir atteindre nos objectifs avant la fin de la quatrième année de doctorat, nous avons commencé en décembre 2004 la phase d'implémentation du prototype. A cet effet, nous avons identifié

le langage de programmation à utiliser. En effet, le projet JXTA est la plate-forme basée sur les technologies Java proposée par Sun Microsystems offrant un ensemble de mécanismes simples permettant le développement d'applications basées sur le Peer-To-Peer. En outre, nous avons commencé à écrire le code source général d'un Peer avec JXTA. Ainsi, la structure de JXTA peut se diviser en différentes couches.

- Le noyau. Il contient les principales fonctionnalités du système, à savoir la gestion des canaux de communication, de la sécurité ou encore des groupes de Peers.
- La couche des services. Certains services sont déjà implantés dans la plate-forme JXTA : ce sont les services de Sun, à savoir l'indexation, la recherche et le partage de données.
- La couche des applications.

Enfin, dans l'étape d'implémentation du prototype, il nous reste à définir nos protocoles et nos fonctionnalités et concevoir l'interface.

Au mois de mars 2005, nous prévoyons commencer nos premières expérimentations. Pour ce faire, il faut étudier les différentes approches possibles. En d'autres termes, il faut identifier les critères les plus appropriés d'évaluation pour déterminer la qualité des documents et la performance du système. À la fin du mois d'avril, nous prévoyons collecter assez d'information pour commencer d'analyse des résultats. Les mois de mai et de juin 2005 seront consacrés à cette dernière étape.

8. Conclusion.

Compte tenu de l'état actuel de notre recherche et du domaine pluridisciplinaire de celle-ci, les contributions scientifiques que nous pouvons escompter de la réalisation du projet proposé sont au nombre de cinq. En premier lieu, l'identification des problèmes spécifiques aux systèmes courants de filtrage d'information. Ensuite, des avancées significatives dans la mise en oeuvre de profils adaptatifs d'utilisateurs, qui mettent à profit les informations issues de l'observation des comportements de l'utilisateur. En outre, une étude sur les modèles sociocognitifs dans les systèmes multi-agents. En plus, la mise en oeuvre d'une technologie « Push » souple et adaptée et qui permet un meilleur recouvrement de l'espace d'information. Enfin, une mise au point de méthodes d'expérimentation et d'évaluation.

Nos futurs travaux prendront plusieurs directions. Nous pouvons classer ces derniers en deux catégories. D'une part, les perspectives à court terme dans lesquelles nous prévoyons de réaliser des expérimentations plus exhaustives. En effet, nous comptons réaliser des expérimentations avec un

grand nombre d'utilisateurs et dans des situations de filtrage d'informations réelles. En plus, nous prévoyons d'identifier d'autres mesures pour l'évaluation de la performance, la confiance et la qualité des documents.

En ce qui concerne nos perspectives à long terme, nous devons prendre en considération la protection des renseignements privés. En effet, dans les réseaux Peer-To-Peer, c'est un sujet de recherche très étudié. Cheval de bataille de Canny (Canny, 2002), les techniques de protection de la vie privée affectent radicalement les performances du système (Olsson, 2003). À cet effet, il faut trouver un compromis entre la sécurité et la performance du système. Nous pouvons penser à l'utilisation des surnoms pour garder l'anonymat et de la cryptographie pour le transfert de données. Dans le même ordre d'idées, nous devons penser à un système d'identification unique de l'utilisateur. En effet, la notion confiance ne peut pas être utilisée si un utilisateur puisse changer d'identification à chaque fois qu'il le veut.

Un autre point intéressant, c'est le partage de la puissance de calcul. En effet, nous pouvons imaginer que si un agent ne possède pas assez de ressources (mémoires, CPU), il peut demander des recommandations à un autre agent dans lequel il a confiance. Enfin, un autre grand point, c'est l'intégration des comportements utilisateurs dans le calcul du degré d'intérêt d'une page (Chan, 1999). En effet, nous pouvons introduire les comportements produits par l'utilisateur en consultant une page (imprimer, défiler, enregistrer, etc.) comme paramètre dans nos calcul du degré d'intérêt d'une page.

Annexes

Annexe 1 : Le cadre de Basili

Définition				
Motivation	Objet	objectif	Domaine	utilisateurs
Comprendre Évaluer Contrôler Apprendre améliorer valider	Produit Processus Modèle Métrique théorie	Caractériser Évaluer Prévoir motiver		
Planification				
Étapes du projet		Intrants		Livrables
Exécution				
Étape 1	Étape 2	Étape 3		Analyse
Interprétation				
Contexte d'interprétation		Extrapolation des résultats		Travaux futurs

Tableau 6 : L'adaptation du cadre de Basili à la recherche exploratoire (Abran et al., 1999).

Bibliographie

Aas, K et Eikvil, L. Text categorisation : a survey. report, 1999.

Abdul-Rahman, A. et S. Hailes , "Supporting Trust in Virtual Communities, "*Proc. 33rd Ann. Hawaii Int'l Conf. System Sciences* (HICSS 33), vol. 6, IEEE CS Press, 2000;

Abran, A, Laframboise, L, et Bourque, P. 1999. "*A Risk Assessment Method and Grid for Software Measurement Programs*". Submitted to Communications of the ACM, Janvier 1999.

Almgren, M. Communicating Agents Developed with Genetic Programming, in: *Genetic Algorithms and GeneticProgramming at Stanford*, June, 2000.

Andrieu, O. 1996. "Méthodes et outils de recherche sur l'Internet". Eyrolles.

Bawa, M. Manku, G.S. et Raghavan, P. 2003. SETS: Search Enhanced by Topic-Segmentation. In Proc. of the 26th Intl. ACM Conf. on Research and Development in Information Retrieval (SIGIR), 2003.

Belew, R, et Van Rijsbergen, C. J. 2001. "Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW ". Cambridge Univ Pr.

Belkin N.J., Croft W.B., « Information filtering and information retrieval: two sides of the same coin? », *Communications of the ACM*, vol. 35, n° 12, p. 29-38, décembre 1992.

Boesch C. et Boesch H., 1989, « Hunting behaviour of wild chimpanzees in the Taï National Park », *American Journal of Physical Anthropology* n° 78, pp. 547-573.

Breese. J, Heckerman. D, et Kadie, k. Empirical analysis of predictive algorithms for collaborative filtering. uncertainty in Artificial Intelligence, 1998.

Byrne, R. W. 1995. The Thinking Ape: Evolutionary Origins of Intelligence.

Oxford University Press.

Burke, R. 2002. Recommender Systems: Survey and Experiments. User Modeling and User-Adapted Interaction 12, 4. 331-370

Byrne R. W. et Whiten A., 1988, Machiavellian Intelligence: social expertise and the evolution of intellect in monkeys, apes and humans, Clarendon Press, Oxford.

Canny, J. 2002. Collaborative Filtering with Privacy, IEEE Conf. on Security and Privacy, Oakland CA, May.

Castelfranchi C. 1990. «Social Power: A Point Missed in Multi-Agent DAI and HCI ».

Chan. P. 1999. A non-invasive learning approach to building user profiles. *Web Usage Analysis and User Profiling*.

Chartron, Ghislaine. 1996. "Recherche d'information sur Internet". ADBS éditions, 1996. p. 43-102.

Cheney D. L. et Seyfarth R. M., 1980, « Vocal recognition in free-ranging vervet monkeys », *Animal Behavior* n° 28, pp. 362-367.

Crespo, A. et Garcia-Molina, H. 2002. "Routing Indices For Peer-to-Peer Systems". Proc. of ICDCS'02, Vienna, Austria.

Croft W.B. 1993. « Knowledge-based and Statistical approaches to Text Retrieval », *IEEE EXPERT*, vol. 8, n° 2, p. 8-12, avril.

Dasser V., 1988, « A Social Concept in Java Monkeys », *Animal Behavior* n° 36, pp. 225-230.

Derudet, G. 1997. "La révolution des agents intelligents". *Internet professionnel*, mai 1997, n. 9, p. 74-80.

De Waal, F. B., 1982, *Chimpanzee Politics : power & sex among apes*, Harper & Row, London.

Drogoul A. et Ferber J., 1992a, « Multi-Agent Simulation as a Tool for Modeling Societies: Application to Social Differentiation in Ant Colonies ».

DROGOUL A. et FERBER J. 1992b. *From Tom Thumb to the Dockers: Some Experiments with Foraging Robots.* In Jean-Arcady Meyer, Herbert Roitblat et Stewart Wilson (éd.) *From Animals To Animats: Second Conference on Simulation of Adaptive Behavior (SAB 92)*, M.I.T. Press, Hawaii.

DROGOUL A. et FERBER J. 1994. *Multi-agent simulation as a tool for studying emergent processes in societies*. In Nigel Gilbert et Jim Doran (éd.) *Simulating Societies: the computer simulation of social phenomena*, Vol. , pp 127-142. Londres, UCL Press.

Drogoul A. et Collinot A., 1997, « Entre réductionnisme méthodologique et stratégie intentionnelle, l'éthologie, un modèle alternatif pour l'IAD ? ». in Quinqueton J., Thomas M.-C., Trousse B. 1997. *Intelligence Artificielle et Systèmes Multi-Agents. Actes des JFIADSMA'97*, Hermès, Paris.

Ferber, J., 1995, "Les systèmes multi-agents", InterEditions.

Finin, T. 1996a. "Agent Related Topics Natural Language Processing and Information Retrieval". URL: www.cs.umbc.edu/agents/related/ir/pns.shtml

Finin, Tim. 1996b. "Agents on, by and for the Web". URL: www.cs.umbc.edu/agents/web/

Foll, Laurent. 1996. La technologie des Info-agents dans l'entreprise et sur Internet. ADBS éditions, p. 121-124.

Foner, L. 1999. "What's an Agent?". URL: foner.www.media.mit.edu/people /foner/ agents.html

Foster, I. Kesselman, C. et Tuecke, S. The anatomy of the Grid: Enabling scalable virtual organizations. Intl. Journal of High Performance Computing Applications, 15(3):200-222.

Franklin, Stan. 1997. "It is an Agent or just a Program? : A Taxonomy for Autonomous Agents". URL : www.msci.memphis.edu/~franklin/AgentProg.html

Frascaria, K. 2002. L2T Bleu, solution de peer-to-peer hybride pour le travail "collaboratif". ZDNet France. Février.

Galliers J. R., 1990, « The Positive Roles of Conflicts in Cooperative Multi-Agent Systems », *in* DEMAZEAU Y. et MÜLLER J.-P. (eds), 1990, *Decentralized AI*, Elsevier (North-Holland), Amsterdam.

Gambetta, D. 2000. Can We Trust Trust? (in Making and Breaking Cooperative Relations), chapter 13, pages 213-237.

Gralla, P, Ishida, S, Reimer, M, et Adams, steph. 1999. "How the Internet Works: Millennium Edition". Que.

Haskin, D. 1997. "WebTamer 1.0: Webtamer: Smooth Surfing". Computer Shopper, 31 Decembre 1997.

Kalogeraki, V. Gunopulos, D. et Zeinalipour-Yazti, D. 2002. "A Local Search Mechanism for Peer-to-Peer Networks". Proc. of CIKM'02, McLean VA, USA.

Lardy, Jean-Pierre. 1996a. “Les outils de recherche d’informations sur Internet : guides, listes thématiques et index ”. *Documentaliste - Sciences de l’information*, janvier 1996, vol.33, n. 1, p. 33-38.

Lardy, Jean-Pierre. 1996b. “Recherche d’information dans Internet : outils et méthodes”. ADBS éditions, 1996.

Lawrence, Steve, et Giles, C. Lee. 1999. “Accessibility of information on the web”. NATURE, vol. 400.

Lieberman, H., Fry, C. et Weitzman, L. 2001. Exploring the Web with Reconnaissance Agents. *Communications of the ACM*, 44(8), 69-75.

Lv, Q. Cao, P. Cohen, E. Li, K. et Shenker, S. 2002. "Search and replication in unstructured peer-to-peer networks". Proc. of ICS02, New York, USA, June.

Mataric, M. J. 1994. Learning to Behave Socially in *Proceedings, From Animals to Animats 3, Third International Conference on Simulation of Adaptive Behavior (SAB-94*).

Mataric, M.J. 1997. Learning Social Behavior, *Robotics and Autonomous Systems*, 1997.

Merkow, M. 2001. Magi At The Heart Of P2P Success. CCP, CISSP. Février.

Minsky M. 1986. « The Society of Mind », Simon and Schuster, New York.

Morikawa, H. Nakauchi, K. Ishikawa Y. et Aoyama, T. 2003. Peer-to-peer keyword search using keyword relationship. In Proceedings of 3rd International Workshop on Global and Peer-to-Peer Computing on Large Scale Distributed Systems (GP2PC), pages 359–366, Tokyo, May.

Müller, J-P. 1998. « Vers une méthodologie de conception de systèmes multi-agents de résolution de problème par émergence », JFIADSMA'98, Pont-à-Mousson, Hermès.

Olsson, T. 2003. Bootstrapping and Decentralizing Recommender Systems. Licentiate thesis 2003-006, Department of Information Technology, Uppsala University.

Oram, A. 2001. Peer-to-Peer: Harnessing the Power of Disruptive Technologies. O'Reilly & Associates, Inc.

Pazzani, M. 1997. A framework for collaborative, content-based and demographic filtering. *Machine Learning*.

Piechaczyk, B. 1996. “Les agents intelligents et les services d'information personnalisée sur Internet”. Institut d'Etudes Politiques, 1996.

Prax, J-Y. 1998. “La gestion électronique documentaire”. InterEditions.

Renaud Chavanne, L. 1997. “Reuters se lance dans le push media pour entreprises”. *Internet Professionnel*, juillet 1997, n. 11, p. 46-47.

Resnick, P. et Varian, H. R. (Ed.). 1997. *Special Section: Recommender Systems*. Communications of the ACM, 40(3).

Roebuck, M. 2000. “Beginner's Guide to Search Engine Placement and Ranking”.

Russell, S.J. et Norvig, Peter. 1995. Artificial Intelligence: A Modern Approach, Englewood Cliffs, NJ: Prentice Hall.

Sarwar, B., Karypis, G., Konstan, J., et Riedl, J. 2000. *Analysis of Recommendation Algorithms for E-Commerce*. In Proceedings of ACM Conference on Electronic Commerce, Minneapolis, MN.

Schapire, Robert. Singer, Yoram et Singhal, Amit. 1998. Boosting and rocchio applied to text filtering. *Research and Development in Information Retrieval*.

Sekaran, U. 1992. Research Methods for Business - A Skill Building Approach, 2nd Edition, Wiley, New York.

Stanley, Tracey. 1997. “Intelligent Searching Agents on the Web”. *Ariadne,* janvier 1997.

Steels, L., 2000. ‘Language as a complex adaptive system’, Lecture notes in Computer Science. Parallel Problem Solving from Nature – PPSN-VI. Schoenaur *et al*, Springer Verlag (Berlin).

Steels, L., 2003. ‘Intelligence with representation’, *Phil. Trans. R. Soc. London.* A (2003) 361, 2381-2395.

Steels, L., 2003. ‘Evolving grounded communication for robots’, Trends in Cognitive Sciences, 7(7), July.

Tsoumakos D. et Roussopoulos, N. 2003. "Adaptive Probabilistic Search for Peer-to-Peer Networks". Proc. of P2P2003, Link□opings, Sweden.

Vaario, J. 1994. « Introduction to Artificial Life », In Robomec'94 Tutorial: Bionic Systems and Artificial Life, Japanese Society of Machinery.

Victor R. Basili, Richard Selby et David Hutchens. 1986. *Experimentation in Software Engineering*. IEEE Transactions on Software Engineering, July.

Widyantoro, Dwi, Ioerger, Thomas, et Yen, John. 2000. Learning user interest dynamics with a three-descriptor representation. *American Society of Information Science*.

Wittner, O. Heegaard, P. E. et Helvik, B. E. 2002. “Swarm based distributed search in the amigos environment,” AVANTEL Technical Report ISSN 1503-4097, Department of Telematics, Norwegian University of Science and Technology, December.

Wooldridge, Michael, et Jennings, Nicholas. 1995. “Intelligent Agents: Theories, Architectures, and Language ”. Springer-Verlag.

Yang, B. et Garcia-Molina, H. 2001. "Comparing hybrid peer-to-peer systems". Proc. of VLDB'01, Rome, Italy.

Yang, B. et Garcia-Molina, H. 2002. "Efficient Search in Peer-to-Peer Networks". Proc. of ICDCS'02, Vienna, Austria.

Yezdi Lashkari. 1999. Feature-guided automated collaborative filtering. *Recommender Systems*.