Vers une approche comportementale de recommandation : apport de l'analyse des usages dans un processus de personnalisation

Ilham Esslimani

Nancy-Université
Université Nancy 2

UFR mathématiques et informatique **École doctorale IAEM Lorraine**

Département de formation doctorale en informatique

Vers une approche comportementale de recommandation : apport de l'analyse des usages dans un processus de personnalisation

THÈSE

présentée et soutenue publiquement le 11 décembre 2010

pour l'obtention du

Doctorat de l'université Nancy 2

(spécialité informatique)

par

Ilham Esslimani

Composition du jury

Rapporteurs :	Pr. Cécile Paris, CSIRO ICT Centre, Australie Pr. Sylvie Calabretto, LIRIS INSA-Lyon
Examinateurs :	Pr. Monique Grandbastien, UHP-Nancy 1 Dr. Jean Philippe Blanchard, Crédit Agricole S.A, Paris
Directrice de thèse :	Pr. Anne Boyer, Université Nancy 2

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Loria

Mis en page avec la classe thloria.

Remerciements

Je tiens à adresser tout d'abord mes remerciements à ma Directrice de thèse Anne Boyer pour son encadrement et ses conseils pendant ces années de thèse. Sa disponibilité, son soutien et son esprit pédagogique m'ont permis d'apprendre beaucoup de choses et de donner le meilleur de moi-même. En outre, sa constante bonne humeur a rendu très agréable nos échanges tout au long de la thèse.
Je remercie également Armelle Brun pour tout le temps qu'elle m'a consacré, pour son esprit d'écoute, pour les échanges intéressants qu'on a eu pendant la thèse et pour tous les conseils qu'elle m'a prodigué. Qu'elle trouve ici l'expression de ma reconnaissance.

Je tiens à exprimer ma gratitude au Groupe Crédit Agricole (S.A) pour avoir soutenu financièrement cette thèse et remercier en particulier Jean Philippe Blanchard pour sa collaboration et pour ses conseils avisés qui m'ont permis de mener à bien mon travail de thèse.

J'adresse mes remerciements également aux membres du jury Cécile Paris et Sylvie Calabretto pour avoir accepté d'être les rapporteurs de ma thèse, Monique Grandbastien et Jean Philippe Blanchard d'avoir été examinateurs de ma thèse.

Je remercie la société Sailendra et en particulier Régis Lhoste pour son assistance, son soutien et sa collaboration.
Mes remerciements vont aussi à tous les membres de l'équipe KIWI que j'ai cotoyés au quotidien. J'ai beaucoup apprécié l'ambiance de travail et les moments agréables passés avec eux qui étaient riches tant sur le plan professionnel que personnel. Je remercie en outre toute l'équipe MAIA de m'avoir accueilli pendant ma première année de thèse.
Mes remerciements s'adressent également à Antoinette Courrier pour son aide notamment pour toutes les procédures administratives qui étaient liées à ma thèse.

Je remercie toute ma famille : mes parents, mes sœurs et mes frères qui m'ont poussé jusqu'au bout pour effectuer cette thèse. Je remercie infiniment mon mari pour son encouragement, son écoute et son soutien tout au long de ces années et grâce à qui j'ai pu surmonter des moments difficiles.
Une pensée très particulière est adressée à Najet Boughanmi, Maha Idrissi Aouad, Geoffray Bonnin, Ahmad Hamad, Chérif Haydar et Rokia Bendaoud. Je remercie aussi tous les amis et les collègues que j'ai côtoyés pendant les années de thèse : Wahiba Touali, Ghaith Kaabi, Hanen Maghrebi, Ines Sakly, Stéphane Goria, Manel Sorba, Ilyess Ohayon, Maxime Rio, Mathieu Lefort, Nicolas Jones, Sylvain Castagnos, Cédric Bernier, Billel Nefzi, Karim Dahman, Yoann Bertrand, Cédric Rose, Walid Fdhila et Arnaud Glad. La liste n'étant pas exhaustive, mes remerciements les plus sincères sont adressés à toute personne que j'ai oubliée de citer ici et qui a contribué de près ou de loin à la réalisation de cette thèse.

Je dédie cette thèse à la mémoire de mon père.

Table des matières

Introduction générale

1 Contexte

Internet est un réseau numérique mettant à la disposition des utilisateurs, notamment à travers le Web et les portails Extranet, une large variété de ressources, appelées aussi "items" qui ont la particularité d'être hétérogènes et distribués et dont le volume est sans cesse croissant. Nous entendons par item tout type de document électronique regroupant un ensemble de données informatives accessible sous un format électronique donné (e.g. format textuel ou multimédia).
Selon une évaluation de l'Internet World Stats[1] réalisée en 2010, il y aurait plus de 1.9 milliards d'internautes dans le monde pouvant consulter environ 109.5 millions de sites Web opérationnels et 25.21 milliards de pages[2]. Or, devant cette surabondance d'items, l'utilisateur devient incapable de gérer cette masse d'information et de repérer les items qui correspondent au mieux à ses attentes, que j'appellerai items pertinents.

Dans ce contexte, le recours à des outils permettant de faciliter l'accès aux items pertinents s'avère crucial. Les moteurs de recherche font partie des premiers outils qui ont été développés pour pallier ce problème d'accès aux items pertinents sur le Web. Ces moteurs ont pour rôle d'explorer et de parcourir le Web afin d'indexer les items qui y sont publiés. Cette indexation consiste en l'extraction de mots-clés, considérés comme significatifs, représentant le contenu des items. L'objectif de ces moteurs de recherche est de proposer des items correspondant aux équations de recherche formulées par les utilisateurs (sous forme de mots-clés).

La dernière décennie a été marquée par une évolution considérable des moteurs de recherche dont *Google* est devenu le plus populaire. Un utilisateur, qui sait *a priori* comment exprimer son équation de recherche, est souvent satisfait par les résultats proposés par un tel moteur de recherche. Cependant, un utilisateur précisant mal ses équations de recherche parce qu'il est peu initié à Internet ou parce qu'il a peu de connaissances sur le sujet recherché, trouvera des difficultés à repérer les items qui correspondent à ses besoins. Ainsi, en choisissant un mot-clé générique tel que "réseau", les moteurs de re-

[1] http ://www.internetworldstats.com
[2] Alessio Signorini. "Indexable Web Size". http ://www.cs.uiowa.edu/∼asignori/web-size/

cherche proposent des milliers voire des millions de résultats se rapportant à différentes thématiques telles que "réseau informatique", "réseau de transport", "réseau d'entreprises" ou même "réseau de trafic de drogue".
De ce fait, la qualité et la pertinence des items proposés par les moteurs de recherche sont notamment conditionnées par la précision des équations de recherche des utilisateurs.

En outre, les techniques utilisées par les moteurs de recherche tel que Google, exploitent principalement le contenu des pages Web ainsi que la structure des hyperliens entre ces pages afin d'évaluer la pertinence et l'importance d'un item par rapport à l'équation de recherche formulée [Brin et Page, 1998]. Peu importe qui a réalisé cette recherche, si la même requête est formulée par deux utilisateurs, les items proposés seront souvent les mêmes. Or, même si deux utilisateurs expriment la même requête, ils n'ont pas nécessairement les mêmes besoins.

Avec l'expansion du Web et le développement de nombreux outils de recherche et de diffusion de l'information, tel que les portails Extranet d'entreprise, l'enjeu est de considérer l'utilisateur lors du processus de recherche d'information [Tamine-Lechani et Calabretto, 2008], en vue de satisfaire ses besoins spécifiques et de le fidéliser ainsi au service en question. Dans le cadre d'un portail Extranet, les utilisateurs étant connus au préalable et non occasionnels, il s'agit de leur facilier l'accès à des informations susceptibles de les intéresser, pouvant être cruciales et nécessaires à l'aboutissement des projets de l'entreprise et contribuant à la prise de décision.

Ces enjeux liés à la satisfaction des attentes des utilisateurs et à leur fidélisation constituent les objectifs principaux de la personnalisation de l'accès à l'information. En effet, la personnalisation a pour finalité de proposer des items en lien avec les goûts réels de chaque utilisateur. La personnalisation est un axe de recherche qui a suscité l'intérêt et l'engouement de nombreux chercheurs. Plusieurs approches ont été ainsi proposées, intégrant les approches basées sur le contenu [Krulwich et Burkey, 1996] [Mladenic, 1999], les techniques à base de critiques issue du domaine de raisonnement à partir des cas ("Case Based Reasoning" (CBR)) [Burke, 2000] [Aha *et al.*, 2000], les approches basées sur la navigation sociale [Svensson *et al.*, 2005], etc.

Les systèmes de recommandation s'inscrivent dans le cadre de la personnalisation de l'accès à l'information. Ils peuvent exploiter les approches citées ci-dessus, en vue de proposer à un utilisateur actif (i.e. un utilisateur courant), des conseils d'items qu'ils jugent pertinents par rapport à ses attentes. Ils cherchent en effet à anticiper ses futurs besoins à travers la prédiction de ses appréciations concernant un ou plusieurs items qu'il n'a pas encore consultés.
En d'autres termes, les systèmes de recommandation ont pour but d'assister l'activité de recherche de l'utilisateur et de l'orienter vers l'information qui lui convient. En guise d'exemple, sur un portail Extranet d'entreprise, le système de recommandation peut proposer à l'utilisateur actif un article spécialisé, une actualité ou bien un rapport technique. Sur un site d'e-commerce, le système de recommandation peut proposer à cet utilisateur un produit à acheter, un livre à lire ou un film à regarder.

Plusieurs techniques, issues notamment du domaine de l'apprentissage automatique et du data mining sont utilisées par les systèmes de recommandations. Le Filtrage Collaboratif (FC) [Goldberg *et al.*, 1992] représente l'une des techniques de recommandation les plus populaires [Adomavicius et Tuzhilin, 2005]. Lorsqu'un utilisateur actif a besoin d'une recommandation, le système de FC retrouve les utilisateurs ayant des préférences et des goûts similaires à cet utilisateur (ces utilisateurs sont appelés "utilisateurs voisins") et utilise leurs opinions pour générer une ou des recommandations susceptibles de l'intéresser.

Dans un processus de recommandation, l'identification des appréciations des utilisateurs est souvent fondamentale, dans la mesure où elle permet de connaître l'utilisateur afin de lui proposer des recommandations pertinentes. Les appréciations reflètent les avis positifs ou négatifs des utilisateurs vis-à-vis d'un certain nombre d'items. Leur identification peut varier selon le type de l'approche utilisée. Par exemple dans un système de recommandation à base de critiques, elle se base sur l'implication directe de l'utilisateur pour l'expression des appréciations, appellée aussi "élicitation". Certes, l'élicitation constitue une démarche fastidieuse pour cet utilisateur [McGinty et Smyth, 2005], puisqu'il est sollicité afin d'exprimer explicitement l'intérêt qu'il porte à un certain nombre d'items. De ce fait, le recours à l'élicitation doit dépendre de l'enjeu de l'approche utilisée.
En effet, dans le cas où cette élicitation va à l'encontre des priorités de l'approche de recommandation, en provoquant par exemple la démotivation et l'abandon de l'utilisateur, le recours à d'autres méthodes d'identification des appréciations s'avère indispensable. Dans cette optique, l'approche par l'analyse des usages peut se présenter comme une solution palliant ce problème.
L'intérêt de cette approche est d'éviter l'élicitation en observant le comportement de l'utilisateur actif et en analysant ses actions lors de son interaction avec un système informatique tel qu'un portail Extranet. L'analyse des usages est ainsi susceptible de ressortir des indicateurs permettant de déduire les appréciations de cet utilisateur et d'identifier éventuellement des communautés virtuelles.

Dans le cadre de cette thèse, nous nous intéressons à l'étude des systèmes de recommandation fondés sur le filtrage collaboratif exploitant l'analyse des usages dans le contexte d'un Extranet d'entreprise. La section qui suit présente les questions de recherche que nous traitons à travers cette thèse.

2 Problématique de recherche

Comme nous l'avons indiqué précédemment, les systèmes de recommandation visent à personnaliser l'accès à l'information fournie par un système informatique. Pour atteindre cet objectif, les systèmes de recommandation peuvent notamment exploiter la technique du FC afin de modéliser les utilisateurs et leur recommander des items pertinents en se

basant sur les opinions de leurs voisins (cf. section 1). Différentes questions de recherche peuvent ressortir de cette définition :

1. **En terme de modélisation des utilisateurs**. Afin de construire un modèle de l'utilisateur actif, le système a besoin notamment de collecter les données relatives aux appréciations de cet utilisateur. L'analyse de ces données permet ensuite de construire ce modèle utilisateur qui va être utilisé par le système pour recommander les items estimés pertinents pour cet utilisateur.
   De ce fait, l'exploitation des appréciations dans un tel processus de recommandation est primordiale. Or, souvent les données relatives aux appréciations ne sont pas suffisamment disponibles dans le système voire pas disponibles du tout [Sarwar *et al.*, 2000b]. Par conséquent, quand le système manque de données, la modélisation des utilisateurs devient difficile et complexe. En effet, dans le cadre du FC, le système serait incapable d'identifier un nombre significatif de voisins nécessaires au calcul de recommandations adaptées aux besoins de l'utilisateur actif.
   En outre, l'enjeu quant à l'exploitation des données d'appréciation est que, du point de vue utilisateur, les contraintes liées à leur collecte doivent être faibles. Il s'agit d'éviter l'intervention directe de l'utilisateur (l'élicitation) pour exprimer ses appréciations parce que d'une part, l'utilisateur dispose de peu de connaissances sur les items pour pouvoir les évaluer tous, et d'autre part, parce qu'il a tendance à être réticent quant à l'évaluation d'items [Burke, 2002].

2. **En terme d'identification de voisins pertinents**. Les systèmes de recommandation à base de FC peuvent utiliser l'approche "kNN" (k Nearest Neighbors) [Resnick *et al.*, 1994], qui repose sur la recherche des "plus proches voisins", afin de calculer les recommandations. L'identification des plus proches voisins consiste à sélectionner les k voisins les plus similaires à l'utilisateur actif. Pour l'évaluation des similarités, cette approche prend en considération les appréciations relatives aux items communs à l'utilisateur actif et les autres utilisateurs. Néanmoins, un système basé sur une approche kNN peut être confronté à une situation où les utilisateurs n'ont pas d'items communs avec l'utilisateur actif (donc pas de voisins). Ainsi, faute de voisins, il sera incapable de proposer des recommandations à cet utilisateur. A cet effet, l'utilisation d'autres techniques permettant de découvrir les similarités entre utilisateurs s'avère cruciale.

3. **En terme de recommandation de la nouveauté**. Lorsqu'un nouvel item est introduit dans le système, il ne peut pas être pris en compte dans le cadre de recommandations basées sur le FC, étant donné que les appréciations des utilisateurs vis-à-vis de cet item ne sont pas encore disponibles. Ce problème est connu sous le nom de "démarrage à froid" ou de "latence" [Schein *et al.*, 2002]. Les systèmes de recommandation doivent ainsi faire face à ce problème dans le but de prendre en considération les nouveaux items au niveau des recommandations proposées à l'utilisateur.

4. **En terme de précision des recommandations** [Herlocker *et al.*, 1999]. Cette question est étroitement liée aux deux premières questions de recherche citées ci-dessus. En effet, la précision des recommandations fournies par un système de recommandation dépend essentiellement de la disponibilité des données permettant de modéliser les utilisateurs et d'identifier des voisins pertinents et fiables. En outre, la performance du système en terme de précision ou qualité de recommandation, émane également de la fiabilité de l'algorithme de modélisation utilisé.
   A cet effet, pour atteindre une meilleure performance en terme de précision, les systèmes de recommandation ont pour enjeu de fournir à l'utilisateur actif des recommandations fiables correspondant à ses besoins, ce qui permettra de le fidéliser le plus possible et d'améliorer l'usage du système informatique en question.

5. **En terme de réduction du temps de calcul et de l'espace de recherche**. La performance d'un système de recommandation est évaluée également au niveau du temps de calcul. En effet, le temps de traitement requis pour le calcul des recommandations doit être réduit, notamment par la réduction de l'espace de recherche utilisé au niveau de la modélisation. Cet enjeu est lié également au passage à l'échelle, lorsque le système dispose d'un nombre considérable d'utilisateurs et d'items à traiter. D'autant plus, ce nombre évolue dynamiquement dans le temps, d'où l'intérêt de la réduction de l'espace de recherche dans le processus de recommandation.

6. **En terme de robustesse**. Le système de recommandation doit être robuste pour faire face aux données bruitées et garantir la fiabilité des recommandations.

La problématique scientifique que nous traitons est liée à la modélisation des utilisateurs en se basant sur l'observation du comportement et sur l'analyse des usages dans le cadre d'un processus de recommandation exploitant le filtrage collaboratif. Notre objectif est de remédier au problème de manque de données, de démarrage à froid et d'améliorer la précision des recommandations. En outre, il s'agit de garantir la robustesse du système de recommandation.

3 Approche proposée

3.1 Cadre industriel

Cette thèse s'inscrit dans le cadre du projet PERCAL réalisé en collaboration avec le *Crédit Agricole S.A*, en particulier avec le Pôle Innovation qui est chargé de l'étude, de l'expérimentation et de la définition des modalités de mise en œuvre des technologies au service des métiers bancaires au sein du Groupe Crédit Agricole.
A partir des questions de recherche soulevées et en prenant en compte le contexte d'un

Extranet d'entreprise, l'objectif de ce projet est de proposer de nouvelles techniques de recommandation permettant l'accès personnalisé à l'information, afin d'optimiser l'usage des ressources de l'Extranet documentaire par les utilisateurs du Groupe Crédit Agricole. En effet, les items et les utilisateurs de cet Extranet étant très nombreux et variés (des milliers d'utilisateurs et des dizaines de milliers d'items), l'enjeu est de pouvoir mettre en place des outils de recommandation collaboratifs, s'appuyant sur l'analyse des usages, capables de mettre à la disposition des utilisateurs des informations pertinentes adaptées à leurs besoins.

3.2 Approche

L'objectif de cette thèse est d'utiliser l'approche par analyse des usages afin de construire des modèles utilisateurs à partir de l'observation de leur comportement navigationnel. En effet, notre hypothèse est que l'analyse des traces d'usage, qui représentent l'ensemble des actions et des événements résultant du processus d'interaction d'un utilisateur avec le système, peut extraire un certain nombre d'indicateurs reflétant les appréciations de cet utilisateur.
Analyser les usages va permettre ainsi de cerner le comportement de l'utilisateur, de connaître mieux ses besoins, ce qui permettra d'améliorer potentiellement les performances et la qualité des recommandations calculées par le système de recommandation. En outre, étant donné que la quantité de traces et d'observations à traiter par le système de recommandation est importante, notre objectif consiste également à proposer une approche permettant de réduire l'espace de recherche lors de l'apprentissage des modèles utilisateurs et pour la génération des recommandations.
De plus, cette approche de recommandation doit permettre de faire face au problème de manque de données. A ce niveau, notre hypothèse est que les techniques issues du domaine de l'analyse des réseaux sociaux peuvent être des solutions prometteuses face à ce problème de manque de données grâce à la découverte de nouvelles relations entre utilisateurs.

3.3 Contributions

Les contributions de cette thèse comprennent :

- Un modèle de recommandation basé sur le filtrage collaboratif comportemental [Esslimani *et al.*, 2008b] [Esslimani *et al.*, 2008a]. Ce modèle exploite les observations relatives au comportement de navigation des utilisateurs pour les modéliser et se base sur le FC pour produire des recommandations. Ce modèle vise à améliorer la qualité des prédictions et à garantir la robustesse du système de recommandation.

- Un modèle de recommandation combinant le filtrage collaboratif comportemental

avec une approche de clustering calculant les clusters selon les similarités de voisins entre utilisateurs [Esslimani *et al.*, 2009a]. Ce modèle a pour objectif de réduire l'espace de recherche des voisins et d'améliorer le temps de calcul des recommandations ainsi que leur précision.

– Un modèle de recommandation exploitant les méthodes de prédiction de lien dans un réseau comportemental [Esslimani *et al.*, 2009b] [Esslimani *et al.*, 2009c] [Esslimani *et al.*, 2010a]. Dans l'objectif d'améliorer l'identification des voisins dans le cadre de ce réseau, ce modèle utilise les associations transitives et les méthodes de prédiction de lien afin d'établir de nouvelles relations entre utilisateurs. Ce modèle a pour enjeu de faire face au problème de manque de données et d'améliorer la précision des recommandations.

– Un modèle de recommandation basé sur les leaders comportementaux pour la recommandation de la nouveauté [Esslimani *et al.*, 2010c] [Esslimani *et al.*, 2010b]. Ce modèle vise à détecter des leaders dans l'objectif de remédier au problème de démarrage à froid dans le cadre d'un réseau comportemental. Ces leaders ont la particularité d'être au "centre" de ce réseau et disposent d'une potentialité importante de prédiction des appréciations des autres utilisateurs concernant les nouveaux items introduits dans le système.

3.4 Evaluation

Pour la validation des approches proposées dans cette thèse, nous avons évalué les différents modèles au travers d'expérimentations sur un corpus d'usage réel qui contient les traces d'usage extraites de l'Extranet du Crédit Agricole. De plus, nous avons utilisé le corpus Movielens (corpus de référence dans le domaine des systèmes de recommandation) du laboratoire de recherche Grouplens[3] afin de confronter certains de nos résultats avec ceux de la communauté scientifique.
Ces approches ont été évaluées en termes de précision, de temps de calcul et de robustesse et comparées au FC standard [Herlocker *et al.*, 1999] utilisé souvent dans les travaux de recherche comme banc d'essai ("benchmark").

Les résultats de ces expérimentations ont été publiés dans :
– des revues internationales : Journal of Digital Information Management (JDIM) [Esslimani *et al.*, 2008a], the Social Network Analysis and Mining Journal (SNAMJ) [Esslimani *et al.*, 2010a] ;
– des conférences internationales : WEBIST 2009 [Esslimani *et al.*, 2009a], ASONAM 2009 [Esslimani *et al.*, 2009b], EC-WEB 2010 [Esslimani *et al.*, 2010c], ASONAM 2010 [Esslimani *et al.*, 2010b] ;
– un workshop international : RSPR 2008 [Esslimani *et al.*, 2008b] ;

[3]http ://www.grouplens.org

– un colloque francophone : ISKO 2009 [Esslimani *et al.*, 2009c].

4 Structure du document

Dans ce manuscrit, nous présenterons dans la première partie le contexte général en décrivant l'origine des systèmes de recommandation ainsi que les données exploitables dans le cadre des recommandations. De plus, il sera question de décrire les principales techniques de recommandation en discutant leurs avantages et leurs inconvénients tout en soulignant les verrous scientifiques que nous traitons dans le cadre de cette thèse (cf. chapitre 1, partie 1).
En outre, dans le chapitre suivant (cf. chapitre 2, partie 1), nous introduirons le schéma générique de la recommandation, tel que nous le percevons. Ensuite, il s'agira de décrire le contexte applicatif lié à nos travaux de recherche ainsi que la méthodologie expérimentale (corpus et mesures d'évaluation) que nous avons utilisée en vue d'évaluer la performance de nos approches.

Les parties suivantes sont consacrées à la description de nos contributions.
La deuxième partie comprend la présentation de l'approche collaborative comportementale de recommandation. Ainsi, nous décrirons dans un premier temps (cf. chapitre 1, partie 2) notre modèle fondé sur le filtrage collaboratif comportemental et les résultats de son évaluation. Ensuite, nous montrerons l'apport d'une approche de clustering exploitant les voisinages dans le cadre du filtrage collaboratif comportemental, notamment en terme de qualité de recommandation (cf. chapitre 2, partie 2).

La troisième partie est dédiée à la description de l'approche sociale de recommandation. Il s'agit de discuter d'abord l'intérêt de faire appel aux méthodes de prédiction de lien dans le cadre d'un réseau comportemental, afin de pallier le problème de manque de données. Dans la même perspective, il est question d'introduire la détection de leaders dans le cadre de ce réseau, pour la recommandation de la nouveauté. Cette partie intègre également les expérimentations qui ont été réalisées pour valider nos modèles et mettre en évidence leur performance, comparés à des modèles de l'état de l'art.

La dernière partie de la thèse comprend la conclusion et les perspectives de recherche. Cette partie résume les principales contributions de la thèse et présente quelques orientations futures de nos travaux de recherche dans le cadre des systèmes de recommandation.

Première partie

Contexte

Chapitre 1

Etat de l'art

Ce chapitre a pour objectif de faire un tour d'horizon, non exhaustif, des systèmes de recommandation liés au domaine de la recherche d'information, en évoquant leur origine et leurs applications et en décrivant les données qu'ils exploitent. De plus, il s'agit de présenter les principales techniques de recommandation en soulignant leurs apports et leurs limites et de discuter les principaux verrous scientifiques auxquels nous nous intéressons dans le cadre de cette thèse.

1.1 Origines et applications

Les systèmes de recommandation ont été utilisés afin de faire face au problème de surcharge et de profusion d'informations disponibles notamment à travers le Web ou les e-services. Les systèmes de recommandation visent à proposer à un utilisateur actif une ou des recommandations d'items susceptibles de l'intéresser. Ces recommandations peuvent concerner un article à lire, un livre à commander, un film à regarder, un restaurant à choisir, etc.

"Tapestry" [Goldberg *et al.*, 1992] représente l'un des premiers systèmes de recommandation. Il a été développé en 1992 par le centre de recherche de "Xerox" aux Etats Unis. Il s'agit d'un système de recommandation intégré à une application de mail électronique, permettant de recommander des listes de diffusion aux utilisateurs. Tapestry est fondé sur le Filtrage Collaboratif (FC) exploitant les annotations (les tags) des utilisateurs attribués aux listes de diffusion. L'analyse de ces annotations par le système de FC permet de déterminer et de proposer les listes de diffusion qui sont pertinentes pour chaque utilisateur. Par la suite, d'autres systèmes de recommandation ont vu le jour en 1994 et en 1995, tels que le système de recommandation d'articles d'actualités et de films développé par "GroupLens" [Resnick *et al.*, 1994] et le système de recommandation de musique "Ringo"

proposé par [Shardanand et Maes, 1995]. Ces deux systèmes sont également basés sur le FC.

Quelques années plus tard, avec l'essor de l'Internet et des applications Web, il y a eu un engouement pour les systèmes de recommandation qui se sont développés dans différents domaines d'applications. Nous pouvons en citer :

- les systèmes de recommandation de films, tels que : Movielens[4] [Herlocker *et al.*, 1999] et Eachmovie [Breese *et al.*, 1998],
- les systèmes de recommandation de livres (Bookcrossing[5] [Ziegler *et al.*, 2005]),
- les systèmes de recommandation de musique (LastFM[6] [Jäschke *et al.*, 2007]),
- les systèmes de recommandation d'articles d'actualités [Billsus *et al.*, 2002],
- les systèmes de recommandation de blagues (Jester[7] [Goldberg *et al.*, 2001]),
- les systèmes de recommandations introduits sur des sites e-commerce (Amazon[8] [Linden *et al.*, 2003]),
- les systèmes de recommandation de restaurants [Burke, 2002],
- les systèmes de recommandation intégrés aux Extranets documentaires (l'Extranet documentaire du Crédit Agricole [Bertrand-Pierron, 2006]),
- les systèmes de recommandations intégrés aux moteurs de recherche (le moteur de recherche d'AOL[9] [Pass *et al.*, 2006]),
- les systèmes de recommandations implémentés sur des sites de recrutement (JobFinder [Rafter *et al.*, 2000]),
- les systèmes de recommandations de citations bibliographiques [McNee *et al.*, 2002] [Cosley *et al.*, 2002].

Pour tous les systèmes de recommandation développés jusqu'à nos jours, la collecte de données relatives aux utilisateurs et/ou aux items, représente une phase clé dans le processus de personnalisation. La section qui suit décrit en détails la typologie de données exploitables par les systèmes de recommandation ainsi que les enjeux liés à leur collecte.

1.2 Données

Dans le cadre des systèmes de recommandation exploitant notamment le FC, la détermination des appréciations est requise afin de pouvoir modéliser l'utilisateur. Cette démarche d'identification d'appréciations repose soit sur des approches dites "réactives" ou soit dites "proactives" [Anand et Mobasher, 2005]. Dans le cas d'une approche réactive, l'utilisateur réagit suite à la demande du système afin d'exprimer ses besoins, tandis que

[4]http ://www.grouplens.org
[5]http ://www.informatik.uni-freiburg.de/∼cziegler/BX
[6]http ://www.lastfm.fr
[7]http ://eigentaste.berkeley.edu
[8]http ://www.amazon.com
[9]http ://www.gregsadetsky.com/aol-data

dans une approche proactive, l'utilisateur est moins sollicité, c'est le système qui anticipe ses besoins.

Dans les approches réactives, la personnalisation est considérée comme un processus conversationnel fondé sur des interactions explicites avec l'utilisateur dans l'objectif d'affiner ses appréciations. Ce processus est réalisé via un ensemble de questions nécessitant un retour de l'utilisateur qui doit exprimer explicitement ses appréciations concernant des critères ou des items.

Les systèmes de recommandation de type réactif, utilisent pour la plupart, des techniques à base de critiques, issues du raisonnement à partir des cas [Smyth, 2007]. L'élicitation du retour de l'utilisateur y est un composant principal permettant d'adapter précisément les recommandations aux besoins exprimés par cet utilisateur.
Par exemple, "Entree" [Burke, 2000] est un système de recommandation de restaurants réactif qui utilise des requêtes, à partir desquelles l'utilisateur spécifie le type de cuisine, le prix, le style de restaurant, la localité, l'atmosphère, etc. L'utilisateur peut ainsi soit accepter les recommandations proposées ou bien les critiquer à travers des critères spécifiques (moins cher, plus calme, etc.).
D'autres exemples de système à base de critique sont proposés également par [Aha *et al.*, 2000], [Shimazu, 2001] et [McGinty et Smyth, 2005].

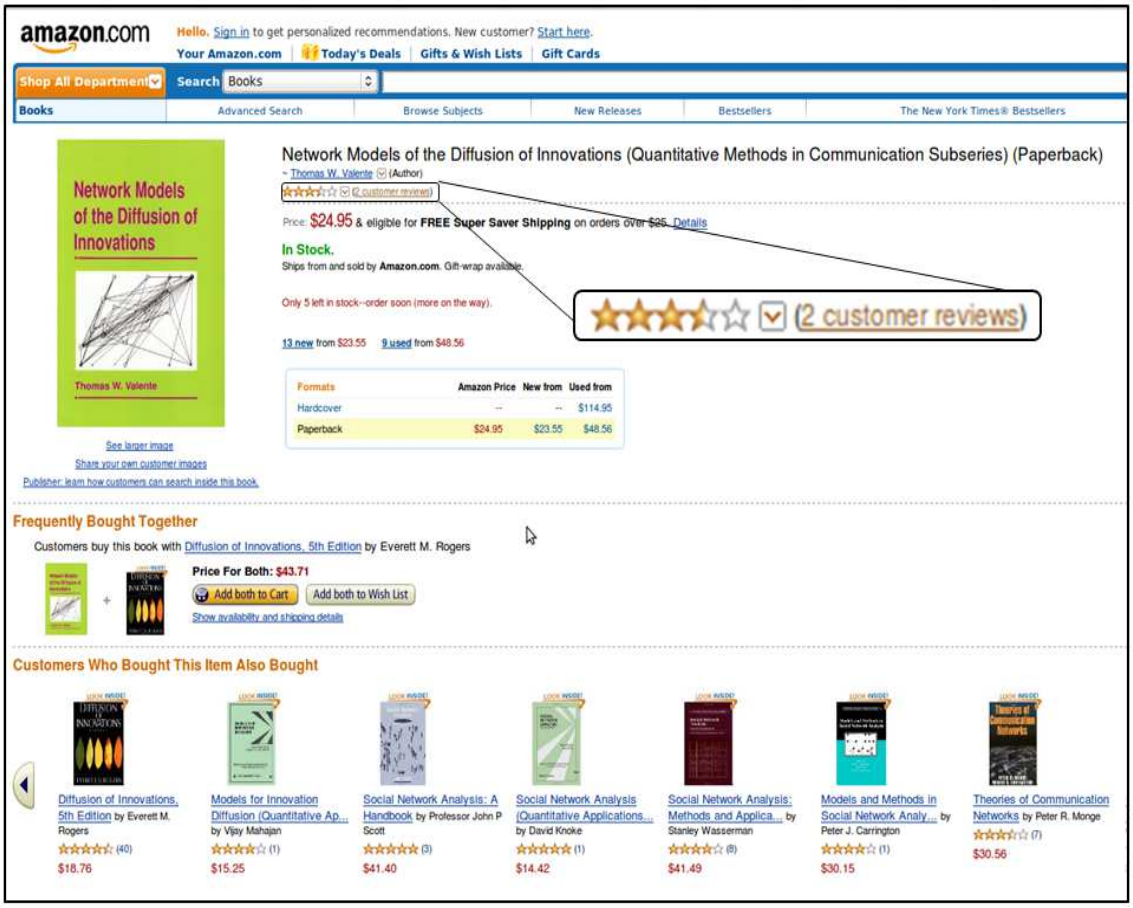
L'avantage des systèmes à base de critique est qu'ils sont faciles à appliquer et ne requièrent pas une connaissance approfondie du domaine de la part de l'utilisateur. Toutefois, les critiques demeurent une arme à double tranchant. En effet, si elles représentent des informations explicites sur les appréciations, elles nécessitent un effort et un investissement de l'utilisateur quant à l'expression de ses avis et de ses retours [McGinty et Smyth, 2005].

Les approches proactives privilégient plutôt la déduction des appréciations pour fournir des recommandations. Les systèmes de recommandation proactifs ne nécessitent pas de retour de l'utilisateur (suite aux recommandations) afin d'orienter le processus de recommandation. Ces systèmes reposent sur l'observation des interactions de l'utilisateur afin d'estimer ses goûts.
Cette observation peut être directe ou indirecte. Quand elle est directe, elle se base sur des données exprimées explicitement par l'utilisateur en attribuant par exemple :

1. des notes aux items consultés indiquant le degré d'appréciation d'un item par cet utilisateur. Les notes sont souvent numériques et limitées par une échelle de valeurs. Une note (numérique) élevée signifie que l'utilisateur accorde un grand intérêt à l'item et qu'il correspond bien à ses goûts. Cependant, une note faible signifie que l'utilisateur ne s'intéresse pas à l'item. Dans d'autres cas, les notes peuvent être exprimées sous une forme binaire telle que "Aime" ou "Aime pas".
   La Figure 1.1 présente un exemple tiré du site de vente en ligne "Amazon" qui offre la possibilité de noter des items (par exemple le livre "Network models of the diffusion

of innovation") sur une échelle de $[1-5]$.

FIG. 1.1 – Exemple de notes : Site d'Amazon



D'une manière générale, l'échelle de note doit refléter les appréciations d'un utilisateur vis-à-vis d'items. Les échelles de note les plus communes sont présentées dans le tableau 1.1 [Schafer *et al.*, 2007]. Le choix d'une échelle de note très large telle que $[1-100]$ peut augmenter l'incertitude sur la valeur de note attribuée. Ainsi, il est difficile de déterminer par exemple la différence entre une note de 55 et de 60 sur l'échelle $[1-100]$, l'écart étant difficilement interprétable par le système et la nuance difficile à évaluer pour un utilisateur.

TAB. 1.1 – Les échelles de notes les plus communes

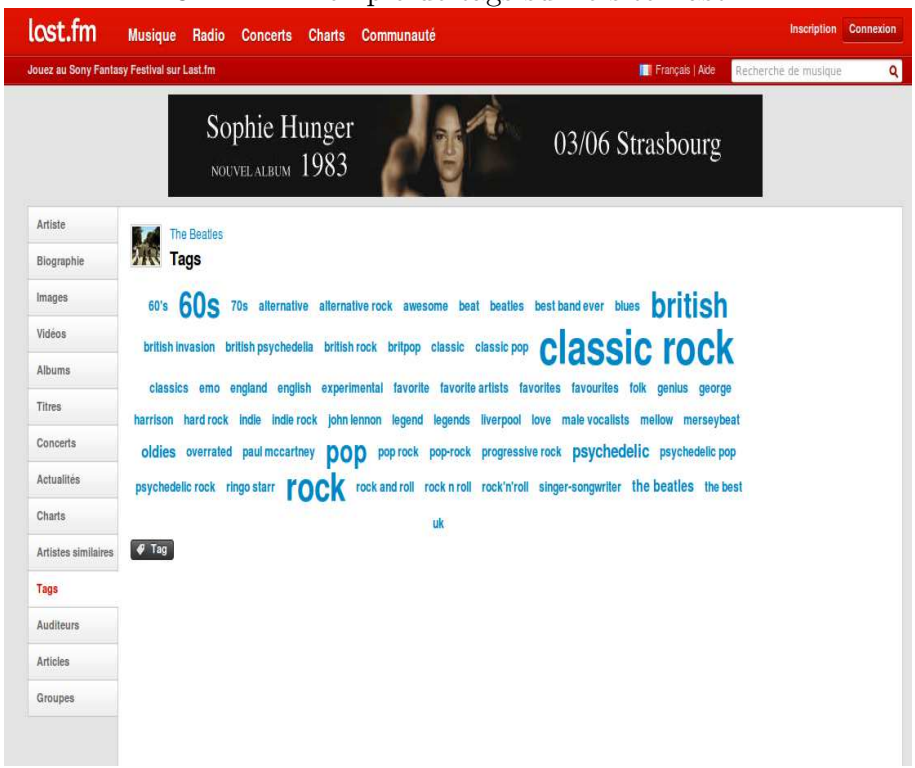
Echelle de note	**Description**
Unaire	"Aime" ou "Je ne sais pas"
Binaire	"Aime" ou "Aime pas"
Entier	$[1-5]$, $[1-7]$ ou $[1-10]$

L'utilisation des notes permet de faciliter l'apprentissage des appréciations, vu que les notes sont faciles à traiter par le système de recommandation. Néanmoins, dans

certains cas, les utilisateurs n'ayant pas les mêmes façons de noter, les notes peuvent ne pas être fiables. En effet, certains utilisateurs attribuent des notes élevées et d'autres non. Par exemple, sur une échelle $[1-5]$, une note qui vaut 3 peut être négative pour un utilisateur et plutôt neutre pour un autre.

2. des commentaires, des mots-clés ou des tags sur des items. La figure 1.2 montre un exemple d'ajout de tags sur le site de recommandation de musique "LastFM". Ces tags sont exprimés dans un langage libre propre à chaque utilisateur, exprimant le mieux son avis.
   Toutefois, tout comme les systèmes à base de critiques, l'expression des appréciations via des commentaires ou tags nécessite une motivation de la part de l'utilisateur, puisqu'elle requiert un effort "cognitif" plus important, par rapport à l'attribution des notes. De plus, le traitement de ces commentaires (exprimés en langage libre) par le système de recommandation demeure assez complexe. Le système doit en effet procéder à une analyse du contenu et à une interprétation des commentaires afin d'estimer les appréciations.

FIG. 1.2 – Exemple de tags sur le site LastFM



3. des attributs démographiques concernant l'utilisateur, tels que : l'âge, le sexe, la catégorie socio-professionnelle, le niveau d'étude, la localité géographique, le statut personnel, etc. Certes, ces attributs ne fournissent pas d'informations sur les appréciations, mais ils permettent notamment d'affiner le profil utilisateur afin d'y adapter les recommandations. Ces attributs peuvent être soit renseignés par l'utilisateur lui-même [Krulwich, 1997], ou bien extraits par exemple à partir des pages Web personnelles [Pazzani, 1999]. Par ailleurs, les profils démographiques peuvent

être utilisés pour calculer les recommandations lorsqu'il s'agit de nouveaux utilisateurs [Vozalis et Margaritis, 2006] [Nguyen *et al.*, 2006]. Ainsi, le système de recommandation peut considérer par exemple que les utilisateurs appartenant à des classes démographiques homogènes, ont des goûts similaires et peut exploiter ces similarités pour la génération de recommandations.

Les appréciations explicitement exprimées par l'utilisateur s'inscrivent dans le cadre d'un processus d'élicitation. Malgré son intérêt, l'élicitation présente certains risques [Rashid *et al.*, 2008]. Elle peut être en effet perçue comme un processus long et fastidieux [Burke, 2002], qui requiert un effort de la part de l'utilisateur, ce qui peut engendrer un abandon du processus d'élicitation.
L'enjeu quant à l'acquisition de toutes ces données explicites décrites ci-dessus, est de trouver le compromis entre collecte de données relatives aux appréciations et réduction de l'élicitation.

Quand l'observation des interactions de l'utilisateur est indirecte, elle repose sur des données ou des appréciations "implicites" déduites à partir des actions réalisées par cet utilisateur. Nous appellerons ces actions "les traces d'usage". Ces traces peuvent inclure [Claypool *et al.*, 2001] :

1. Des indicateurs décrivant la manipulation tels que : des "copier/coller" d'un texte à partir d'une page, la recherche d'un texte dans une page, l'ajout ou la suppression d'un item du panier ou la commande d'un item (dans le cadre des applications e-commerce), la sauvegarde ou l'impression d'une page, l'ajout d'une page aux favoris, l'envoi d'une page à un ami, etc.
2. Des indicateurs de navigation tels que : la fréquence et la durée de consultation, le nombre de clics et de survols de souris sur une page et sur des liens, le "scrolling", etc.
3. Des indicateurs externes marquant l'intérêt. Ces indicateurs décrivent les conditions physiques et émotionnelles qui caractérisent un utilisateur lors de son interaction. Ils peuvent être mesurés par exemple par l'oculométrie[10] ("eye-tracking").

L'enjeu quant à la collecte des traces d'usage est de définir des heuristiques afin de déterminer quelles actions ou quelles traces reflètent une appréciation positive ou bien négative. Par exemple, l'action de suppression d'un item d'un panier (sur un site d'e-commerce) peut être interprétée comme un avis négatif. De même, le critère de temps de consultation peut être aussi considéré. Or, le problème qui se pose est de déterminer s'il s'agit réellement d'une consultation de l'item. Il est possible en effet que l'item soit actif pendant une certaine durée, alors que l'utilisateur ne le consulte pas réellement.

Par ailleurs, la démarche de collecte des données explicites ou implicites (les traces

[10]Technique de suivi et d'enregistrement du mouvement oculaire sur un site Web par exemple, pour détecter les zones du site les plus visées par l'utilisateur

d'usage) dans le cadre d'un système de recommandation doit veiller à la préservation de la vie privée et des données personnelles des utilisateurs. En outre, quel que soit le type de données collectées par le système, cette démarche doit prendre en considération la gestion de l'accroissement du volume de données dans le temps.

Dans cette section, nous avons présenté les différents types de données exploitables par les systèmes de recommandation, en discutant leurs avantages et leurs inconvénients. Nous pouvons déduire à partir de ces discussions que la mise en place d'un système de recommandation de type proactif ou réactif, exploitant des observations directes ou indirectes, requiert une réflexion approfondie *a priori* sur la collecte des données, avec ou sans la sollicitation directe de l'utilisateur.

Aprés avoir présenté la typologie des données exploitées en entrée par les systèmes de recommandation, dans la section suivante il est question de décrire les principales techniques de recommandation.

1.3 Techniques de recommandation

Il existe une large variété de techniques de recommandation. A travers les travaux de recherche, différentes tentatives de classification des approches ou des techniques ont été réalisées. La classification de ces approches dépend notamment du type de données exploitées et de la méthode d'apprentissage utilisée par le système de recommandation. Dans cette section, en distinguant la technique basée sur le contenu du FC basé sur la mémoire ou sur un modèle [Anand et Mobasher, 2005] [Su et Khoshgoftaar, 2009], nous présentons les principales techniques de recommandation avec leurs apports et leurs limites.

1.3.1 Technique basée sur le contenu

La technique de recommandation basée sur le contenu repose sur l'hypothèse que des items ayant des contenus similaires seront appréciés pareillement [Schafer *et al.*, 2007]. Pour la proposition de recommandations aux utilisateurs, cette technique est fondée sur l'analyse des similarités de contenu entre les items précédemment consultés par les utilisateurs et ceux qui n'ont pas été encore consultés [Burke, 2002].
Ainsi, afin de recommander par exemple des films à un utilisateur, le système analyse les corrélations entre ces films et les films consultés antérieurement par cet utilisateur. Ces corrélations sont évaluées en considérant des attributs comme le titre et le genre. De ce fait, parmi ces films, ceux qui seront recommandés à l'utilisateur, sont les plus similaires (en terme d'attribut) aux films consultés par cet utilisateur [Adomavicius et Tuzhilin, 2005].

Parmi les premiers systèmes de recommandation basés sur le contenu, nous pouvons citer : NewsWeeder [Lang, 1995], Letizia [Lieberman, 1995] et InfoFinder [Krulwich et Burkey, 1996], etc. [Pazzani et Billsus, 2007] présente une synthèse de ces systèmes de recommandation en s'intéressant en particulier à la représentation du contenu et aux algorithmes utilisés pour la construction des profils utilisateurs.

La technique de recommandation basée sur le contenu peut être appliquée à la recommandation de pages Web, de films, d'articles actualités, de restaurants, etc. Si nous prenons l'exemple d'un système de recommandation d'articles scientifiques basé sur le contenu, lorsqu'un utilisateur a tendance à consulter souvent des articles portant sur le domaine de la génétique, le système lui proposera des recommandations liées à la génétique. En effet, ces articles disposent de mots-clés communs tels que : "ADN", "gène" ou "protéine".
Il est à signaler que ces mots-clés sont généralement soit extraits sur la base d'une indexation automatique, soit attribués manuellement.
Pour ce qui est des systèmes de recommandation de films ou de restaurants, le contenu est plutôt structuré et réprésenté par des métadonnées définies au préalable et valables pour tous les items [Pazzani et Billsus, 2007].

Dans le cadre de la technique basée sur le contenu, la mesure TF-IDF ("Term Frequency-Inverse Document Frequency") [Salton, 1989] représente la mesure la plus populaire pour l'analyse du contenu. Il s'agit d'une mesure statistique qui permet d'évaluer l'importance d'un mot dans un document ou dans un item faisant partie d'une collection ou d'un corpus [Pazzani et Billsus, 2007].
Le principe de cette mesure est que les mots-clés paraissant dans beaucoup d'items ne permettent pas de distinguer un item pertinent d'un autre qui ne l'est pas. Or, les mots-clés qui sont rares et communs à quelques items définissent plus la similarité de contenu ainsi que la pertinence d'un item par rapport à un autre.

La technique basée sur le contenu a pour avantage de pouvoir générer des recommandations en dépit d'une situation de démarrage à froid. Le démarrage à froid se traduit notamment par l'introduction d'un nouvel item au système de recommandation. Lorsque ce système exploite le filtrage collaboratif, il ne sera pas capable d'incorporer ce nouvel item aux recommandations, puisque les notes relatives à cet item ne sont pas encore disponibles. Ainsi, grâce à l'analyse de contenu, cet item peut être intégré aux recommandations proposées à un utilisateur actif.
Néanmoins, la technique basée sur le contenu présente quelques limites, notamment :

- Le manque de diversité et la surspécialisation des recommandations. En effet, les items recommandés sont toujours similaires et identiques (en terme de contenu) aux items précédemment consultés par l'utilisateur. Les autres items, ayant un contenu non similaire, ne sont jamais intégrés aux listes de recommandation, alors qu'ils pourraient intéresser l'utilisateur.

- La représentation des items est toujours limitée aux descriptions ou aux attributs

qui leur sont associés. Par conséquent, afin d'avoir un ensemble suffisant d'attributs, il est nécessaire soit de prétraiter le contenu pour permettre une extraction automatique d'attributs, soit d'attribuer les descriptions manuellement [Shardanand et Maes, 1995]. Dans les deux cas, l'extraction d'attributs demeure une opération fastidieuse surtout lorsqu'il s'agit d'items multimédia tels que : les images, les documents audio et vidéo, etc. De ce fait, certains aspects pertinents du contenu peuvent être négligés, ce qui peut avoir un impact sur la qualité des recommandations.

Dans les sections suivantes, nous nous intéressons aux approches qui font abstraction du contenu. Ces approches, basées sur le FC, exploitent notamment les appréciations (explicites et/ou implicites) ainsi que les traces d'usage des utilisateurs dans le cadre des recommandations.
Ces approches reposent en effet sur l'hypothèse que les utilisateurs qui partageaient les mêmes goûts dans le passé (en attribuant des notes similaires, en achetant les mêmes articles ou en visitant les mêmes items), vont très probablement avoir les mêmes goûts dans le futur [Goldberg *et al.*, 2001].

1.3.2 Méthodes basées sur la mémoire

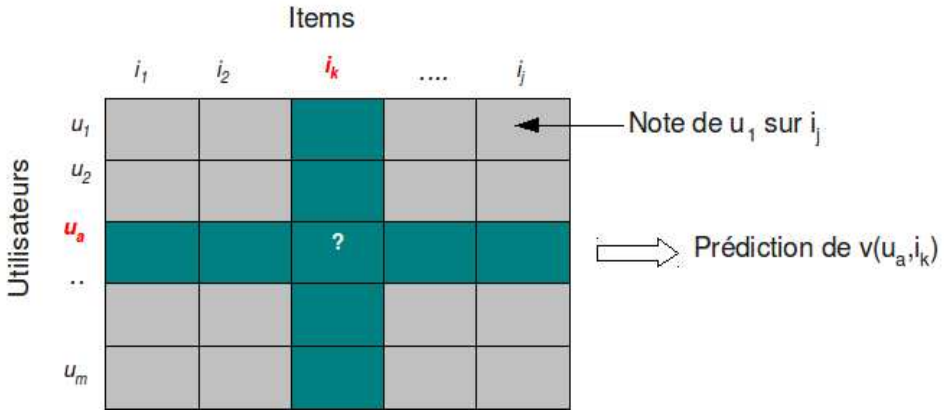
L'approche basée sur la mémoire exploite les appréciations des utilisateurs sur les items (sous forme de notes par exemple), afin de générer les prédictions [Sarwar *et al.*, 2001]. Cette approche applique principalement des techniques statistiques dans le but d'identifier des utilisateurs voisins ayant, sur un même ensemble d'items, des appréciations similaires à celles de l'utilisateur actif. Une fois les voisins identifiés, l'approche basée sur la mémoire utilise différents algorithmes afin de combiner les appréciations des voisins et générer des recommandations à l'utilisateur actif [McLaughlin et Herlocker, 2004].

Dans ce contexte, la technique la plus utilisée et la plus populaire est le *Filtrage Collaboratif* (FC) basé sur la mémoire [Goldberg *et al.*, 1992]. Le FC basé sur la mémoire recherche les "k plus proches voisins" (k Nearest Neighbors "kNN") [Resnick *et al.*, 1994], i.e. les k voisins les plus similaires à l'utilisateur actif, dans le but de générer des recommandations fiables. Ces voisins sont identifiés à partir d'une évaluation de la similarité des appréciations sur les items communs à l'utilisateur actif et les autres utilisateurs.

Dans un système de FC basé sur la mémoire, tel que décrit dans la figure 1.3, les données sont représentées sous forme d'une matrice "Utilisateur x Item" (dont un exemple est présenté dans le tableau 1.2), où les lignes représentent les utilisateurs $U = \{u_1, ...u_m\}$ et les colonnes constituent les items $I = \{i_1, ...i_j\}$. Les utilisateurs fournissent leurs opinions concernant les items sous forme de notes v. Pour un utilisateur actif u_a (par exemple Jean) n'ayant pas exprimé son avis concernant un item i_k (le film "Les visiteurs"), le système recherche les utilisateurs voisins les plus proches notés U_a (parmi Rose, Ryan et Hélène ayant noté le film "Les visiteurs" et qui ont déjà co-noté le film "Pulp Fiction" avec Jean) et utilisent leurs opinions pour prédire la note manquante $v(u_a, i_k)$ (v(Jean , Les visiteurs)).

Ainsi, nous pouvons distinguer deux phases essentielles en FC basé sur la mémoire : la phase d'identification du voisinage et la phase de calcul des prédictions. Les sous-sections qui suivent décrivent chacune de ces deux phases.

FIG. 1.3 – Matrice "Utilisateur x Item"



TAB. 1.2 – Exemple de matrice "Utilisateur x Item"

	Pulp Fiction	Star Gate	Les visiteurs	Scream
Jean	1	5	**?**	3
Rose	4	2	4	?
Eric	3	?	?	5
Ryan	4	?	5	?
Hélène	2	?	4	1

Identification du voisinage

Plusieurs mesures ont été exploitées dans le cadre du FC basé sur la mémoire dans le but d'évaluer les similarités d'appréciations entre utilisateurs et identifier les utilisateurs voisins (les plus proches). Parmi ces mesures nous pouvons citer : le coefficient de corrélation de Pearson [Herlocker *et al.*, 1999], la mesure basée sur le cosinus [Sarwar *et al.*, 2000b], la corrélation de Spearman [Resnick *et al.*, 1994], "Mean squared difference" (qui représente une mesure de dissimilarité) [Shardanand et Maes, 1995], etc.
Les mesures les plus populaires sont le coefficient de corrélation de Pearson et la mesure basée sur le cosinus. Cette popularité est liée à leur contribution à la performance des systèmes de recommandation [Anand et Mobasher, 2005].
Nous décrirons ces deux mesures ci-dessous. Notons que $CorrP(u_a, u_b)$ et $Cos(u_a, u_b)$ désignent les similarités calculées respectivement avec le coefficient de corrélation de Pearson et la mesure basée sur le cosinus, entre deux utilisateurs u_a et u_b. I_a et I_b représentent respectivement l'ensemble des items notés par u_a et u_b. $\overline{v(u_a)}$ représente la moyenne de notes de u_a et $v(u_a, i)$ désigne la note de u_a sur l'item i. I_c désigne les items co-notés (notés en commun) entre l'utilisateur actif u_a et l'utilisateur u_b.

– *Le coefficient de corrélation de Pearson* : cette mesure est présentée dans l'équation (1.1). Lorsque $CorrP(u_a, u_b)$ vaut 1, cela signifie que les utilisateurs u_a et u_b sont fortement corrélés. Or, si $CorrP(u_a, u_b)$ vaut -1, cela implique que u_a et u_b ont des appréciations totalement opposées. Quand cette corrélation vaut 0, aucune relation n'existe entre les deux utilisateurs.

$$CorrP(u_a, u_b) = \frac{\sum_{i \in I_c}(v(u_a, i) - \overline{v(u_a)})(v(u_b, i) - \overline{v(u_b)})}{\sqrt{\sum_{i \in I_c}(v(u_a, i) - \overline{v(u_a)})^2 \sum_{i \in I_c}(v(u_b, i) - \overline{v(u_b)})^2}} \tag{1.1}$$

– *La mesure basée sur le cosinus* : cette mesure est très fréquemment utilisée dans le domaine de la recherche d'information. Dans ce contexte, elle consiste à évaluer la similarité entre deux documents représentés par des vecteurs de fréquences de mots, en calculant le cosinus de l'angle formé par ces deux vecteurs [Salton et McGill, 1983].
En FC, cette mesure peut être adaptée pour l'évaluation de la similarité entre deux utilisateurs u_a et u_b en calculant le cosinus de l'angle entre les vecteurs correspondant à ces deux utilisateurs sur la base de l'équation (1.2) [Breese *et al.*, 1998], en prenant en considération les items co-notés I_c. La valeur calculée par la mesure cosinus est comprise entre 0 et 1.

$$Cos(u_a, u_b) = \frac{\sum_{i \in I_c} v(u_a, i) * v(u_b, i)}{\sqrt{\sum_{i' \in I_a} v(u_a, i')^2 \sum_{i' \in I_b} v(u_b, i')^2}} \tag{1.2}$$

L'inconvénient des deux mesures Pearson et cosinus, est que le calcul des similarités devient non fiable voire impossible, lorsque le système dispose de peu d'items co-notés entre utilisateurs. Afin de pallier ce problème, certaines extensions ont été proposées notamment par [Breese *et al.*, 1998], telle que "La note par défaut " consistant à attribuer une valeur par défaut à une note manquante. Mais l'enjeu à ce niveau est de savoir quelle valeur par défaut choisir (appréciation positive, négative ou bien neutre) et d'évaluer son impact sur le calcul des similarités.
Par ailleurs, en vue d'améliorer la performance des systèmes de recommandation exploitant le FC basé sur la mémoire, [Breese *et al.*, 1998] ont proposé d'utiliser :
– "L'amplification de cas" permettant de transformer les similarités en amplifiant les valeurs proches de 1 et en pénalisant celles qui sont proches de 0, dans le but d'attribuer un poids important aux voisins fortement similaires à l'utilisateur actif.
– "La fréquence inverse utilisateur" inspirée de la méthode IDF ("Inverse Document Frequency"), présentée dans la section 1.3.1. L'hypothèse est que les items appréciés par un grand nombre d'utilisateurs sont moins pertinents pour le calcul des similarités comparés à ceux qui sont appréciés par un nombre restreint d'utilisateurs. Ainsi, chaque note est transformée en la multipliant par la fréquence inverse utilisateur qui est équivalente à $\log \frac{n}{n_{i_k}}$, n étant le nombre total des utilisateurs et n_{i_k} le nombre d'utilisateurs ayant noté i_k.

Calcul des prédictions

Cette deuxième phase, tout comme la première, est d'une importance cruciale dans la mesure où l'objectif de tout système de FC est le calcul des prédictions pour générer des recommandations pertinentes à un utilisateur actif. La méthode la plus utilisée pour le calcul de ces prédictions est la "*somme pondérée*" [Herlocker *et al.*, 1999]. Suivant l'équation (1.3), cette méthode considère les plus proches voisins U_a (corrélés avec l'utilisateur actif) ayant déjà noté l'item i_k, pour calculer la prédiction de la note de u_a sur i_k notée $Pred(u_a, i_k)$. $Sim(u_a, u_b)$ désigne la valeur de similarité entre u_a et un voisin u_b ($u_b \in U_a$) et peut être instanciée par les similarités calculées à partir du coefficient de Pearson ($CorrP(u_a, u_b)$) ou bien à partir de la mesure basée sur le cosinus ($Cos(u_a, u_b)$).

$$Pred(u_a, i_k) = \overline{v(u_a)} + \frac{\sum_{u_b \in U_a} Sim(u_a, u_b) * (v(u_b, i_k) - \overline{v(u_b)})}{\sum_{u_b \in U_a} Sim(u_a, u_b)} \tag{1.3}$$

Le choix des plus proches voisins U_a est déterminant dans la mesure où la performance du système dépend de la qualité des voisins impliqués lors de la génération des prédictions. Différentes stratégies peuvent être prises en compte pour la sélection de ces voisins :

- La détermination d'un seuil de similarité [Breese *et al.*, 1998] [Shardanand et Maes, 1995] : il s'agit de sélectionner les plus proches voisins qui sont corrélés avec l'utilisateur actif à partir d'un seuil de similarité préétabli.

- La sélection de la taille du meilleur voisinage [Herlocker *et al.*, 1999] : cette stratégie permet de sélectionner les voisins les plus proches (20, 50 ou 100 meilleurs voisins par exemple).

- La détermination d'un seuil pour les items co-notés [Viappiani *et al.*, 2006] : cette stratégie consiste à filtrer les plus proches voisins en fonction du nombre d'items co-notés avec l'utilisateur actif.

Au niveau des trois stratégies, les seuils choisis ne doivent pas avoir des valeurs extrêmes (ni trop élevées, ni trop faibles). En effet, par exemple, si la valeur du seuil de similarité est trop faible, cela peut engendrer de mauvaises prédictions quand l'utilisateur actif est corrélé avec de nombreux utilisateurs. De la même façon, si le seuil est très élevé, cela peut affecter la qualité des prédictions et la couverture (la capacité du système à générer des prédictions), quand l'utilisateur actif est faiblement corrélé avec les autres utilisateurs. En effet, dans ce cas, le système ne dispose que de peu de voisins pour pouvoir générer les prédictions.

Une fois les prédictions calculées, le système de FC recommande à l'utilisateur actif les items ayant les valeurs de prédiction les plus élevées.

Par ailleurs, l'approche basée sur la mémoire peut être centrée sur l'item. Cette approche a été proposée par [Sarwar *et al.*, 2001]. Le principe de cette approche consiste à analyser la matrice "Utilisateur x Item" pour identifier des relations entre les items et utiliser ces relations afin de calculer les prédictions. L'hypothèse est que l'utilisateur serait intéressé par des items, similaires aux items qu'il a appréciés auparavant (i.e. similaires en termes de notes attribuées par cet utilisateur).
Pour [Sarwar *et al.*, 2001], dans ce processus, il n'est pas nécessaire d'identifier les voisinages pour les utilisateurs. Par conséquent, un tel système a tendance à calculer plus rapidement les recommandations et permettre ainsi le passage à l'échelle. Les auteurs supposent en effet que le nombre d'items est généralement moins important que le nombre d'utilisateurs.
Cette hypothèse peut être valable pour les applications en e-commerce, où le nombre potentiel des utilisateurs augmente régulièrement, comparé au nombre de produits proposés. Or, dans d'autres contextes, comme dans un portail Extranet (l'Extranet du Crédit Agricole par exemple), ce n'est pas vraiment le cas. En effet, le nombre d'utilisateurs reste relativement stable par rapport au nombre d'items accessibles qui est de plus en plus croissant.

L'approche basée sur la mémoire a pour avantage la simplicité de l'implémentation et de l'intégration des nouvelles données dans le système. Cependant, cette approche a l'inconvénient d'être très dépendante de la quantité de notes des utilisateurs. En effet, si les données s'avèrent rares, il est difficile d'identifier des voisins fiables (à partir des items co-notés) et par conséquent la performance du système décroît.
De plus, dans une situation de démarrage à froid, cette approche est incapable de tenir compte des nouveaux utilisateurs et/ou items, récemment introduits au système. En effet, l'approche basée sur la mémoire nécessite la disponibilité des appréciations concernant ces utilisateurs et/ou ces items pour pouvoir les intégrer parmi les recommandations.
En outre, l'approche basée sur la mémoire reste limitée dans la mesure où elle ne permet pas le passage à l'échelle. En effet, quand le nombre d'utilisateurs et d'items présents dans le système devient important, la génération des recommandations requiert un temps de traitement très élevé.

1.3.3 Méthodes basées sur un modèle

Les méthodes basées sur un modèle ont été intégrées aux systèmes de recommandation pour remédier aux problèmes des méthodes basées sur la mémoire, dont notamment : la non robustesse au manque de données ainsi que le non passage à l'échelle [Sarwar *et al.*, 2000b] [Su et Khoshgoftaar, 2009]. Pour faire face à ces deux problèmes, les méthodes basées sur un modèle utilisent notamment les techniques de réduction de dimensionnalité ou le clustering dans le but d'écarter les utilisateurs ou les items non représentatifs. Ainsi l'espace de représentation utilisateur-item est plus réduit et le taux de données manquantes est moins important comparé à l'espace de représentation original. Les voisins peuvent ainsi être calculés dans cet espace réduit, ce qui permet de garantir le passage à

l'échelle.

Dans le cadre des méthodes basées sur un modèle, le processus de FC consiste à construire des modèles (généralement en hors ligne "off-line") en exploitant les données collectées sur l'utilisateur et/ou sur l'item. Les modèles construits sont par la suite utilisés pour générer les prédictions qui sont proposées à l'utilisateur actif lors de son interaction avec le système.
Le processus de construction du modèle est basé sur les techniques d'apprentissage automatique, telles que : le clustering, les réseaux bayésiens, les arbres de décision, etc. Ces techniques vont être explicitées dans ce qui suit.

Clustering

Un cluster est une collection d'objets qui sont similaires entre eux et dissimilaires aux objets appartenant aux autres clusters [Han et Kamber, 2001]. Dans le cadre du FC, le clustering a pour objectif de créer des clusters homogènes d'utilisateurs ou d'items. Les prédictions sont par la suite calculées en prenant en considération les opinions des utilisateurs (en FC centré sur l'utilisateur) ou les notes des items (en FC centré sur l'item) faisant partie des mêmes clusters.

Les méthodes de clustering les plus exploitées sont les méthodes de partitionnement dont *k-means* [MacQueen, 1967] est la plus populaire.
Dans le cas d'un clustering d'utilisateurs [Kim *et al.*, 2002], k-means consiste à créer k clusters telle que la distance entre utilisateurs intracluster est faible alors que la distance intercluster est forte. En d'autres termes, chaque cluster créé doit comprendre des utilisateurs ayant des appréciations similaires.
L'algorithme (1) [Han et Kamber, 2001] présente les étapes d'un clustering k-means appliqué aux utilisateurs. Cet algorithme consiste à choisir aléatoirement des k centroïdes (des points situés au centre) à partir de l'espace de représentation (i.e. matrice "Utilisateur x Item"). Par la suite, chaque utilisateur est affecté à un cluster, tel que la distance entre cet utilisateur et le centroïde du cluster est faible. Dans une étape suivante, en prenant en compte les utilisateurs qui viennent d'être affectés aux clusters, la position du centroïde de chaque cluster est recalculée. Après la découverte des nouveaux centroïdes, les distances sont à nouveau réévaluées afin de retrouver le cluster auquel chaque utilisateur devrait appartenir. Cette opération est itérée jusqu'à ce que les centroïdes deviennent stables et ne changent plus.

Pour illustrer ces étapes, la figure 1.4 [Han et Kamber, 2001] présente un exemple permettant la génération de trois clusters ($k = 3$) basée sur k-means.
Au début du processus de clustering, trois utilisateurs représentant les centroïdes (représentés par le symbole "+") sont sélectionnés arbitrairement afin de construire trois clusters. Ainsi, dans la phase (a) chaque utilisateur est affecté au cluster le plus proche. La phase (b) représente l'étape de recalcul des positions des centroïdes ainsi que la réaffectation

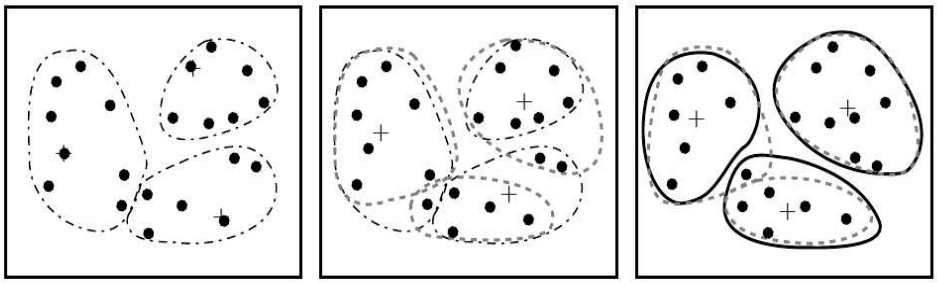
Algorithm 1 Algorithme de partitionnement k-means

1: **Input** : k : le nombre de clusters et M : matrice "Utilisateur x Item"
2: **Output** : k clusters

3: Choisir aléatoirement k centroïdes initiaux de clusters
4: **repeat**
5: Réaffecter chaque utilisateur au cluster auquel il est le plus similaire
6: Recalculer les distances des utilisateurs dans chaque cluster
7: Mettre à jour les centroïdes
8: **until** Stabilité des centroïdes

des utilisateurs aux clusters les plus proches (les différentes lignes pointillées déterminant les trois clusters, changent au fur et à mesure du recalcul des positions des centroïdes). La phase (*c*) représente la fin du processus du clustering, les lignes pleines reflètent les clusters définitifs obtenus suite à la stabilité des centroïdes.

FIG. 1.4 – Clustering k-means



L'algorithme k-means a l'avantage d'être efficient et son implémentation demeure facile [Su et Khoshgoftaar, 2009]. De plus, il permet le passage à l'échelle dans la mesure où il peut être appliqué à de larges corpus. Notons que la complexité de cet algorithme est $O(nkt)$, n étant le nombre total d'utilisateurs, k le nombre de clusters et t le nombre d'itérations.
Toutefois, le choix aléatoire des centroïdes au début du processus du clustering k-means ainsi que la détermination de leur nombre reste encore problématique. [Castagnos, 2008] a étudié ce problème et a proposé d'améliorer le choix des centres initiaux dans le cadre d'un clustering k-means, en garantissant la convergence de l'algorithme lorsque $k = 2$.
Par ailleurs, la méthode k-means demeure sensible aux données aberrantes ("outliers"). Cette sensibilité découle du fait qu'un objet ou un utilisateur ayant une valeur extrêmement différente des autres (un "outlier") peut altérer la distribution de données [Wang et Shao, 2004]. En effet, lorsqu'un outlier est très loin du centroïde d'un cluster, la position du ce centroïde va être déplacée. Par conséquent, la distribution de données ne va plus être homogène.

"PAM" (Partitioning Around Medoïds) est un algorithme de la famille des méthodes de partitionnement. C'est une méthode de clustering de type "k-medoïde" qui a été proposée afin de réduire la sensibilité aux données aberrantes et de remédier au problème de recouvrement des clusters [Han et Kamber, 2001].
Cette méthode de partitionnement a pour objectif de créer un ensemble de clusters tel que chaque cluster ait un point représentatif (un utilisateur central) appelé "médoïde". L'algorithme (2) décrit les étapes du clustering PAM [Han et Kamber, 2001].
Au début du processus, les utilisateurs représentatifs (médoïdes) u_{med} de chaque cluster sont choisis aléatoirement, comme dans k-means. Par la suite, afin d'identifier les médoïdes effectifs, la méthode PAM repose sur la minimisation des dissimilarités entre chaque utilisateur u_p et l'utilisateur représentatif du cluster u_{med}.
L'algorithme PAM itère jusqu'à ce que que les médoïdes deviennent stables, i.e., jusqu'à ce que les u_{med} ne changent plus. Durant cette itération, la qualité du clustering est évaluée en utilisant une fonction qui calcule le coût total S. Ce coût mesure l'erreur en cas de permutation d'un médoïde intial u_{med} avec un autre médoïde u_{random}. Si S est négative, u_{med} est remplacée effectivement par u_{random}. Autrement, u_{med} est considérée comme acceptable et devient stable.

Algorithm 2 Algorithme de partitionnement PAM

Input : k : le nombre de clusters et M : matrice "Utilisateur x Item"
Output : k clusters

Choisir aléatoirement k utilisateurs comme étant les médoïdes initiaux de clusters
repeat
    Affecter chaque utilisateur à un cluster tel que la dissimilarité entre cet utilisateur et le médoïde est faible
    Sélectionner aléatoirement un utilisateur non-représentatif (non-médoïde) u_{random}
    Calculer le coût total, S, de permutation d'un utilisateur représentatif u_{med} avec u_{random}
    if $S < 0$ **then**
        Remplacer u_{med} par u_{random} pour former les nouveaux médoïdes
    end if
until Stabilité des médoïdes

Comme nous l'avons précisé ci-dessus, l'intérêt de l'algorithme PAM comparé à k-means, réside dans son insensibilité aux données aberrantes [Kaufman et Rousseuw, 1990] [Wang et Shao, 2004]. Cette insensibilité est dû au principe même de l'algorithme. En effet, au lieu de considérer une valeur située au centre des utilisateurs comme étant le point de référence dans un cluster (comme dans k-means), PAM désigne des utilisateurs réels représentatifs des clusters (médoïdes) parmi les autres utilisateurs. Un médoïde constitue l'objet ou l'utilisateur le plus central du cluster. Ceci est assuré en permutant systématiquement un médoïde et un autre utilisateur choisi aléatoirement afin de vérifier si la qualité du clustering décroît [Tufféry, 2007].
Néanmoins, l'algorithme PAM reste inapproprié pour de larges corpus. Il requiert en effet, un temps de traitement plus important que l'algorithme k-means. En effet, la complexité

de cet algorithme est $O(tk(n-k)^2)$. De plus, comme k-means, la méthode PAM nécessite également de définir k qui est le nombre de clusters à générer.

Dans le cadre des systèmes de recommandation, la méthode de partitionnement k-means a été largement appliquée aux utilisateurs et/ou aux items, en vue de réduire l'espace de recherche et le temps de calcul des recommandations, de permettre le passage à l'échelle et de pallier le manque de données [Tang et McCalla, 2003] [Xue *et al.*, 2005] [Jiang *et al.*, 2006]. Or, à notre connaissance, la méthode PAM a été moins utilisée par les systèmes de recommandation [Wang *et al.*, 2008].
Par ailleurs, pour ces mêmes perspectives, d'autres algorithmes de clustering ont été intégrés aux systèmes de recommandation, notamment : ROCK [Conner et Herlocker, 1999], Gibbs Sampling [Breese *et al.*, 1998], etc.
Toutefois, l'une des limites du clustering est le risque de perte d'information cruciale lors de la création des clusters. Par exemple, suite à un clustering, deux utilisateurs proches peuvent ne pas avoir été affectés au même cluster, ce qui peut se répércuter sur la performance du système de recommandations.

Modèles probabilistes

Les modèles probabilistes utilisés dans le cadre du FC visent à représenter le calcul des prédictions sous forme de distributions de probabilité [Schafer *et al.*, 2007]. Ces modèles évaluent en général la probabilité qu'un utilisateur u_a attribue une note v à un item i_k, notée $Pr(v(u_a, i_k))$ [Breese *et al.*, 1998]. La note v est comprise entre v_{min} et v_{max} qui représentent respectivement la valeur minimale et maximale correspondant à l'échelle de note. i_x désigne un item appartenant à I_{u_a} qui constitue l'ensemble des items notés par u_a.

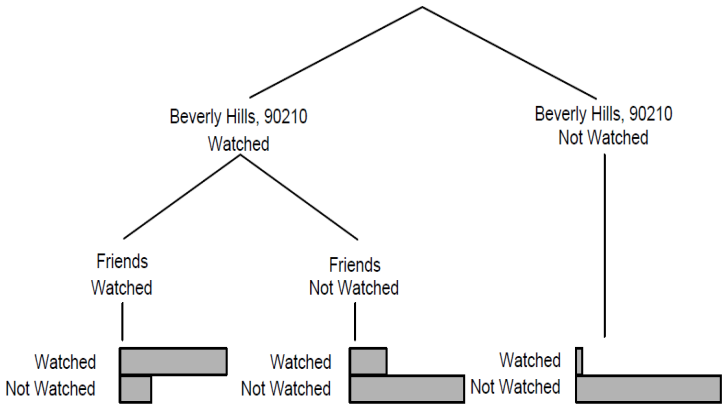
$$Pred(u_a, i_k) = \sum_{v=v_{min}}^{v_{max}} Pr(v(u_a, i_k) = v | v(u_a, i_x), i_x \in I_{u_a}) * v \tag{1.4}$$

Les modèles probabilistes appliqués au FC intègrent notamment les réseaux bayésiens. [Breese *et al.*, 1998] sont parmi les premiers à avoir proposé des méthodes probabilistes pour le FC basé sur les réseaux bayésiens et exploitant les arbres de décision. Le FC est ainsi perçu comme un réseau bayésien où chaque item représente un nœud. Les états de chaque nœud correspondent aux valeurs possibles de note. Ces valeurs comprennent aussi l'état "pas de note" correspondant à une note manquante. Ainsi, pour prédire ces notes manquantes, un algorithme d'apprentissage de réseaux bayésiens est appliqué. Dans les réseaux résultant de cet apprentissage, chaque item dispose d'un item parent à travers un arbre de décision qui définit les probabilités conditionnelles qu'un item soit apprécié ou pas par l'utilisateur.
[Breese *et al.*, 1998] montrent que les réseaux bayésiens exploitant les arbres de décision

améliorent la précision des items recommandés, comparés au FC basé sur l'approche mémoire.

La figure 1.5 présentée par [Breese *et al.*, 1998] est un exemple d'un arbre de décision qui représente les probabilités estimées (représentées par des barres), qu'un utilisateur regarde ou pas la série "Melrose Place", sachant que les nœuds parents sont les séries "Beverly Hills 90210" et "Friends". Par exemple, nous pouvons observer que les utilisateurs n'ayant pas regardé "Beverly Hills 90210", ne vont très probablement pas regarder "Melrose Place".

FIG. 1.5 – Exemple d'arbre de décision présenté par [Breese *et al.*,1998]



Il existe d'autres approches probabilistes, appliquées notamment pour la réduction de la dimensionnalité [Schafer *et al.*, 2007]. Ainsi, une variable dite cachée $Pr(z|u_a)$ est utilisée. Cette variable représente la probabilité qu'un utilisateur u_a appartienne à une classe cachée z. L'équation (1.5) permet de calculer la probabilité qu'un utilisateur u_a attribue une note v à un item i_k.

$$Pr(v|u_a, i_k) = \sum_z Pr(v|i_k, z)Pr(z|u_a) \tag{1.5}$$

Ainsi, la prédiction de v est calculée sur la base de l'équation (1.6).

$$Pred(v|u_a, i_k) = \sum_v (v * \sum_z Pr(v|z, i_k)Pr(z, u_a)) \tag{1.6}$$

Pour l'estimation des classes z, l'algorithme "Expectation-Maximization" peut être appliqué dans le cadre de l'analyse sémantique latente ("Latent Semantic Analysis") [Hofmann, 2004].
Par ailleurs, dans le cadre des modèles probabilistes, d'autres techniques peuvent être également exploitées dans un processus de recommandation, notamment : la décomposition

en valeurs singulières (SVD "Singular Value Decomposition") et l'analyse en composantes principales (PCA "Principal Component Analysis") [Sarwar *et al.*, 2000b] [Goldberg *et al.*, 2001].

Les modèles probabilistes permettent de pallier le problème de manque de données et d'améliorer la qualité des recommandations [Breese *et al.*, 1998]. Néanmoins, la construction des réseaux bayésiens demeure coûteuse et donc inappropriée pour un grand volume de données.

La sous-section suivante est consacrée à la présentation des techniques issues du Web Usage Mining. Ces techniques font partie des méthodes basées sur un modèle, mais au vu de leur importance vis-à-vis de nos travaux de recherche, nous avons choisi de leur consacrer une sous-section à part.

1.3.4 Techniques issues du Web Usage Mining

Le Web Usage Mining (WUM) consiste en l'analyse du comportement de l'utilisateur en se basant sur l'observation et l'analyse de ses activités de navigation et de ses échanges interactifs (ses usages) [Srivastava *et al.*, 2000]. La finalité des techniques du WUM est de pouvoir découvrir des comportements communs d'usage entre utilisateurs afin de générer des prédictions sur les futurs comportements de ces utilisateurs lors de leurs prochaines navigations.

Pour une analyse efficiente des usages, une collecte de traces d'usage est nécessaire. Les traces d'usage représentent une suite d'actions effectuées par un utilisateur, elles sont déduites de l'ensemble des clics effectués par cet utilisateur (cf. section 1.2).

Le WUM est une approche qui occupe une place de plus en plus prépondérante dans plusieurs domaines dont les applications sont relatives notamment aux portails d'information, au e-commerce/e-marketing, au e-learning et à l'IHM (Interaction Homme-Machine), etc. Sur un portail d'information, le WUM permet de prédire quel article sera lu ; sur un site de vente en ligne, il permet de savoir quel produit sera acheté ; et sur un site e-learning, le WUM permet de découvrir par exemple quelles suites d'actions mènent à la réussite ou à l'échec d'un exercice [Cheype, 2006]. De plus, le WUM peut être également utilisé pour améliorer la structure d'un site Web en mettant en évidence des liens hypertextes qui devraient relier des pages Web.

Dans l'objectif de générer des prédictions, le WUM exploite notamment les techniques d'apprentissage automatique pour la découverte des *motifs d'usage* [Srivastava *et al.*, 2000]. Ces motifs permettent de prédire les futurs comportements navigationnels de ces utilisateurs en se basant sur l'analyse de leurs traces d'usage. Ainsi, contrairement aux méthodes de recommandation présentées dans les sections précédentes, les données de notes ne sont pas nécessaires dans le cadre des techniques du WUM.

Les sous-sections suivantes présentent quelques méthodes et algorithmes utilisés dans ce cadre.

Règles d'association

Initialement, les techniques de découverte de règles d'association ont été développées pour l'analyse des bases de données transactionnelles [Agrawal et Srikant, 1994]. Par la suite, ces techniques ont été intégrées dans d'autres domaines, notamment dans le cadre du WUM [Srivastava *et al.*, 2000].
Au niveau d'une base de données transactionnelle, les techniques de découverte de règles d'association permettent la découverte de corrélations entre items. Ces corrélations sont identifiées à travers l'exploration de probabilités estimant que si un certain nombre d'items sont présents, d'autres items sont également potentiellement présents dans la même transaction [Wang et Shao, 2004].

La découverte de règles d'association dans une base de données transactionnelle repose sur deux étapes essentielles :

- La découverte d'itemsets fréquents. Un itemset désigne un ensemble d'items qui apparaissent dans une même transaction. Cette découverte est basée sur "le support" qui détermine la fréquence minimum d'apparition de ces itemsets dans la base de données.
- La découverte des règles d'association à partir des itemsets fréquents en se basant sur "la confiance". La confiance évalue le degré d'implication d'une règle d'association. Si la confiance est élevée, la règle est fiable.

Considérons un ensemble de transactions T intégrant un ensemble d'itemsets $I = \{I_1, I_2, ...I_n\}$. Le support d'un itemset $I_i \subset I$ est définie par l'équation (1.7). $|T|$ représente le cardinal de T.

$$\sigma(I_i) = \frac{|t \in T : I_i \subseteq t|}{|T|} \tag{1.7}$$

Une règle d'association "r" est exprimée sous la forme $X \Rightarrow Y(\sigma_r, \lambda_r)$ [Agrawal *et al.*, 1993] [Anand et Mobasher, 2005]. X et Y représentent des itemsets. $\sigma_r = \sigma(X \cup Y)$ est le support de $X \cup Y$, il représente la probabilité que X et Y se trouvent ensemble dans une transaction.
λ_r est la confiance de la règle r, telle que définie par l'équation (1.8). Cette équation calcule la probabilité que Y apparaisse dans une transaction étant donné que X est déjà apparu dans cette même transaction.

$$\lambda_r = \frac{\sigma(X \cup Y)}{\sigma(X)} \tag{1.8}$$

Dans le cadre du WUM, la découverte des règles d'association est d'un intérêt considérable. Par exemple, pour un Extranet d'entreprise ou un portail d'information, les règles d'association permettent d'observer que les utilisateurs consultant un item i_1, consultent souvent un item i_2. Cette règle aura la forme de $i_1 \Rightarrow i_2$.

Les règles d'association ont été largement utilisées par les systèmes de recommandation [Krulwich, 1997] [Sarwar *et al.*, 2000a] [Fu *et al.*, 2000] [Lin *et al.*, 2002] [Nakagawa et Mobasher, 2003] [Wang et Shao, 2004]. Toutefois, cette technique présente quelques limites. En effet, quand le système manque de données, les règles d'association et les recommandations ne peuvent pas être calculées. De plus, le processus de calcul de règles requiert un temps de calcul élevé et devient non-performant quand la taille de données est importante.

Motifs séquentiels

La recherche de motifs séquentiels, introduite par [Agrawal et Srikant, 1995], peut être considérée comme une variation des règles d'association. En effet, elle repose sur le principe d'ordre des éléments ou de temporalité dans le but de découvrir des séquences fréquentes ordonnées dans le temps [Gery et Haddad, 2003]. A la différence des règles d'association, elle pose plus de contraintes.
Un exemple de motif séquentiel est que les utilisateurs ont tendance à consulter dans l'ordre, sur un portail d'information, les articles : "Volcan d'Islande", puis "Suspension des vols en Europe" et enfin "Prévisions météorologiques".

A l'instar des règles d'association, la recherche de motifs séquentiels a été appliquée d'abord aux bases de données transactionnelles dans le cadre des stratégies marketing [Han et Kamber, 2001]. Ainsi, il était possible d'identifier par exemple que "les clients qui ont acheté l'appareil photo numérique Samsung, vont probablement acheter plus tard une imprimante HP".
Par la suite, d'autres domaines d'applications se sont intéressés à l'étude des séquences de données, telles que :

- Le Web mining qui comprend le WUM et le "Web Structure Mining" (WSM).
  Dans le domaine du WSM [Srivastava *et al.*, 2000], l'étude des séquences vise à analyser la structure de sites Web dans l'objectif d'identifier les liens hypertextes et les pages Web les plus populaires (au travers des usages) et d'en faciliter l'accès.
  Dans le WUM, l'étude des séquences de navigation permet notamment l'aide à la navigation sur le Web [Baumgarten *et al.*, 2000], [Mobasher *et al.*, 2001], [Nakagawa et Mobasher, 2003], [Gery et Haddad, 2003].

– L'analyse des séquences biologiques (séquences ADN ou de protéines) : étude de l'alignement de séquences afin de détecter d'éventuelles anomalies ou disfonctionnements génétiques [Brazma *et al.*, 1998].

– La détection d'intrusions sur des bases de données : mise en place de systèmes de détection de transactions malveillantes grâce aux motifs séquentiels [Hu et Panda, 2004].

Etant donné un ensemble de séquences sur lesquelles les motifs séquentiels seront appris, chaque séquence "s" est représentée par une suite d'événements qui se sont produits l'un après l'autre. En considérant un support minimum, l'analyse de motifs séquentiels permet de retrouver toutes les séquences fréquentes dont la fréquence d'occurences parmi l'ensemble des séquences, est supérieur au support minimum [Agrawal et Srikant, 1995]. Lorsqu'un client réalise par exemple des achats, ces derniers constituent des événements et vont représenter une séquence pour ce client. Un client achète d'abord des items en s_1, puis en s_2, etc. Le nombre d'items dans une séquence représente la longueur de la séquence [Han et Kamber, 2001].

Le tableau 1.3 est un exemple d'une base de données transactionnelle triée par client et par date de transaction. Si nous considérons un support de 25%, $\langle(i_3)(i_9)\rangle$ et $\langle(i_3)(i_4i_7)\rangle$ sont les séquences permettant de satisfaire le support défini et représentent les motifs séquentiels. En effet le motif $\langle(i_3)(i_9)\rangle$ est présent chez les deux clients Jean et Ryan et le motif $\langle(i_3)(i_4i_7)\rangle$ est présent chez Rose et Ryan. Même si Rose a acheté l'item i_6 en même temps que les items i_4 et i_7, $\langle(i_4i_7)\rangle$ représente un motif puisqu'il est une sous-séquence de $\langle(i_3)(i_4i_7)\rangle$.
Il est à signaler que ces motifs sont inter-transactions, alors que les règles d'association sont intra-transactions (i.e. elles sont extraites d'une même transaction).

TAB. 1.3 – Exemple de base de données transactionnelle

Client Id	Date	Items Id
Jean	25 Mai	i_3
Jean	30 Mai	i_9
Rose	10 Mai	i_1,i_2
Rose	15 Mai	i_3
Rose	20 Mai	i_4,i_6,i_7
Eric	25 Mai	i_3,i_5,i_7
Ryan	25 Mai	i_3
Ryan	30 Mai	i_4,i_7
Ryan	25 Juin	i_9
Hélène	12 Mai	i_9

Dans le cadre du WUM, l'analyse des motifs séquentiels peut mettre en évidence des motifs séquentiels de type contigu (fermé) ou bien non contigu (ouvert) [Anand et Mobasher, 2005]. Les motifs contigus sont une forme restrictive des motifs séquentiels. En

effet, la particularité des motifs contigus est que les items contenus dans le motif séquentiel doivent être adjacents suivant l'ordre de la séquence. Par exemple un motif séquentiel contigu $\langle i_4 i_5 i_6 \rangle$ est satisfait par la séquence $\{i_4, i_5, i_6\}$ et non pas par $\{i_4, i_5, i_8, i_6\}$ qui représente plutôt un motif séquentiel ouvert, étant donné que i_8 apparaît entre i_5 et i_6.

L'utilisation des motifs séquentiels pour la recommandation de pages Web est d'un grand intérêt. Cependant, cette technique s'avère limitée lorsqu'il est question de traiter un grand volume de traces d'usage et de générer des motifs en temps réel.

Par ailleurs, il existe d'autres techniques permettant la découverte des motifs séquentiels, telle que :

- La technique LCS (Longuest Common Subsequences) : C'est une technique issue de la programmation dynamique. Elle permet d'extraire un cas particulier de motifs séquentiels. En effet, cette technique vise à identifier la plus longue sous-séquence commune à deux séquences données. Dans le cadre des systèmes de recommandation, [Jalali *et al.*, 2008] ont proposé une architecture de classification des motifs séquentiels, en se basant sur la découverte de LCS. Ces motifs permettent de prédire les futures activités de navigation des utilisateurs. Dans [Banerjee et Ghosh, 2001], un algorithme basé sur la technique LCS est utilisé pour le clustering d'utilisateurs en exploitant les traces d'usage. Cette approche de clustering prend en compte les similarités entre les chemins de navigation, basées sur les LCS, ainsi que la durée de consultation des items contenus dans ces LCS.

- Les modèles de Markov : cette approche vise à mettre en évidence des liens séquentiels entre les items consultés durant les activités de navigation des utilisateurs. En estimant les probabilités conditionnelles de transition entre items, les dépendances séquentielles de comportement de navigation sont modélisées sur la base des modèles de Markov [Eirinaki *et al.*, 2005]. Plusieurs travaux de recherche ont intégré les modèles de Markov dans le processus de recommandation notamment : [Zimdars *et al.*, 2001], [Shani *et al.*, 2005], [Liu *et al.*, 2007], [Bonnin *et al.*, 2009] et [Verma *et al.*, 2009].

Différents algorithmes ont été proposés pour la recherche de motifs séquentiels depuis leur émergence en 1994, notamment : GSP [Srikant et Agrawal, 1996], FreeSpan [Han *et al.*, 2000], SPADE [Zaki, 2001], SPAM [Ayres *et al.*, 2002], etc.
Tous ces algorithmes ont été intégrés dans diverses applications. Dans le domaine du WUM, de nombreux travaux de recherche ont eu un engouement pour les motifs séquentiels, notamment : [Baumgarten *et al.*, 2000], [Gaul et Schmidt-Thieme, 2001], [Mobasher *et al.*, 2001], [Nakagawa et Mobasher, 2003], [Gery et Haddad, 2003].

Les techniques issues du WUM présentées ci-dessus, ont pour avantage d'analyser les usages et de prédire les futurs comportements navigationnels des utilisateurs sans l'utilisation des notes (requises notamment dans l'approche basée sur la mémoire). Or, comme pour les autres approches présentées précédemment, les algorithmes d'extraction

de motifs séquentiels ou de règles d'association traitent les données sans prendre en compte leur évolution dynamique dans le temps. Le passage à l'échelle, l'optimisation du temps de calcul et la génération de motifs en temps réel demeurent encore des enjeux de taille. De plus, dans le cadre de ces techniques, seul le critère de consultation d'items est considéré (pour la recommandation de pages Web par exemple). En effet, deux utilisateurs ayant visité les mêmes items, auront les mêmes recommandations, alors qu'ils peuvent avoir des goûts différents.

1.3.5 Techniques hybrides

Les différentes techniques exploitées par les systèmes de recommandation ont chacune leurs apports mais aussi leurs limites. Le tableau 1.4 présente une synthèse comparant les avantages et les inconvénients des techniques de recommandation qui ont été présentées dans cet état de l'art.
Nous pouvons observer à partir de ce tableau que le FC basé sur un modèle peut être performant, cependant cette performance reste un compromis entre amélioration de la qualité des recommandations et construction coûteuse de modèles.
Quant au FC basé sur la mémoire, bien qu'il soit fiable et simple à implémenter, il demeure peu performant surtout lorsque le système manque de données, telles que les notes.
La technique basée sur le contenu permet de remédier à ce problème de manque de données. Toutefois, les recommandations qu'elle génère sont très spécialisées et manquent de diversité (i.e. les items recommandés à un même utilisateur ont un contenu similaire).
Ainsi, le choix d'une technique de recommandation reste un compromis entre performance, facilité d'implémentation et complexité.

De ce fait, afin de combler les faiblesses d'une technique par une autre, plusieurs travaux de recherche ont proposé de combiner ou d'hybrider des techniques de recommandation qui sont potentiellement complémentaires.

Le système de recommandation hybride le plus courant consiste à combiner les techniques basées sur le contenu avec le FC basé sur la mémoire [Balabanović et Shoham, 1997], [Pazzani, 1999], [Claypool *et al.*, 1999], [Schein *et al.*, 2002]. Il existe différentes possibilités de combinaison, [Adomavicius et Tuzhilin, 2005] les ont classifié en quatre catégories :

- Implémenter séparément le FC basé sur la mémoire et les méthodes basées sur le contenu et combiner les prédictions par la suite en se basant sur une combinaison linéaire des notes prédites.

- Incorporer certaines caractéristiques issues du contenu dans le cadre du FC basé sur la mémoire. De ce fait, au lieu de calculer les similarités sur la base des items co-notés comme en FC, les similarités entre utilisateurs sont évaluées en se basant sur la corrélation du contenu des items consultés [Balabanović et Shoham, 1997].

TAB. 1.4 – Synthèse comparative des techniques de recommandation

Catégorie	**Exemples d'algorithmes utilisés**	**Avantages**	**Inconvénients**
Technique basée sur le contenu	– Analyse de similarité de contenu (TF/IDF) – Clustering – Arbres de décision	– Amélioration de la qualité des recommandations – Réduction du problème de manque de données	– Manque de diversité des recommandations – Nécessité d'indexation de contenus (extraction d'attributs représentatifs) – Problème d'indexation de documents multimédia
FC basé sur la mémoire	– FC exploitant l'approche kNN (basée sur l'utilisateur ou sur l'item) – Utilisation des mesures Pearson ou cosinus	– Implémentation simple – Intégration facile de nouvelles données – Précision des recommandations	– Dépendance aux données de notes – Détérioration de la qualité de recommandations à cause du manque de données – Problème de passage à l'échelle
FC basé sur un modèle	– Clustering – Approches probabilistes (réseaux bayésiens) – Méthodes de réduction de dimensionnalité (SVD, PCA) – WUM (règles d'association, motifs séquentiels, modèles de Markov)	– Amélioration de la qualité des recommandations – Réduction du problème de manque de données – Prédiction des futurs comportements de navigation	– Construction coûteuse de modèles – Risque de perte d'information pertinente dû à la réduction de dimensionnalité – Problème de calcul des règles ou de motifs quand le système manque de données – Pas de considération du profil utilisateur (pour les modèles du WUM)

Cette stratégie de combinaison permet de pallier certains problèmes de manque de données, dûs par exemple à un faible nombre d'items co-notés entre utilisateurs.

- Incorporer certaines caractéristiques issues du FC basé sur la mémoire dans le cadre d'une approche basée sur le contenu. Il s'agit de créer par exemple une vue collaborative des profils utilisateurs qui sont représentés par des vecteurs de termes extraits du contenu des items [Soboroff et Nicholas, 1999].

- Construire un modèle général unifiant les caractéristiques issues à la fois du contenu ainsi que du FC basé sur la mémoire. [Popescul *et al.*, 2001] proposent en effet une méthode probabiliste afin d'unifier ces caractéristiques en se basant sur l'analyse sémantique latente.

[Burke, 2002] analyse également les différentes stratégies de combinaison de techniques de recommandation d'une manière générale. Il présente ainsi différentes méthodes d'hybridation, dont notamment :

- La méthode pondérée : les notes calculées par les différentes techniques de recommandation sont combinées et pondérés afin de générer une seule recommandation. L'intérêt de cette méthode est que la combinaison est simple à réaliser et permet d'ajuster l'hybridation en fonction des performances.

- La méthode "switching" : le système change à chaque fois de technique de recommandation selon les performances atteintes dans le but de ne conserver que les meilleures prédictions. Proposé par [Billsus *et al.*, 2000], le système "DailyLearner" utilise cette méthode dans le cadre d'une hybridation entre contenu et FC basé sur la mémoire. Ce système applique d'abord la méthode basée sur le contenu. Lorsque cette dernière génère des recommandations de faible qualité, le système fait appel à la technique du FC.
  La méthode "switching" introduit une complexité supplémentaire au processus de recommandation. En effet, le critère permettant le choix d'une technique doit être déterminé, ce qui requiert un autre niveau de paramètrage sur le système.

- La méthode mixte : les recommandations issues de différentes techniques sont toutes présentées simultanément aux utilisateurs. Le problème qui peut ressortir quant à l'utilisation de cette méthode est la difficulté de calculer les scores pour ordonner une liste de recommandation, lorsque toutes les techniques recommandent les mêmes items mais avec des notes différentes.

Les différents travaux qui s'intéressent à l'hybridation de techniques de recommandation ont démontré empiriquement que cette hybridation permet d'améliorer la précision des recommandations comparée par exemple au FC basé sur la mémoire ou à la technique basée sur le contenu [Balabanović et Shoham, 1997] [Pazzani, 1999] [Melville *et al.*, 2002]. Les systèmes hybrides permettent en outre de remédier à certains problèmes tels que le manque de données.

Toutefois, l'hybridation rajoute encore plus de complexité au processus de recommandation [Su et Khoshgoftaar, 2009]. En effet, elle requiert différentes sources de données et met en application plusieurs techniques à la fois. Elle nécessite ainsi des paramétrages supplémentaires liés à la combinaison de différentes méthodes. Par conséquent, les calculs requis pour cette hybridation deviennent coûteux.

Après avoir présenté les principales techniques utilisées par les systèmes de recommandation, dans la section suivante il est question de discuter les verrous scientifiques auxquels nous nous intéressons dans cette thèse.

1.4 Verrous scientifiques

Malgré le succès des systèmes de recommandation, certains points demeurent encore problématiques, notamment : le manque de données, le démarrage à froid, la sélection de voisins fiables, la robustesse et la précision des recommandations. Cette section vise à expliciter ces points problématiques en soulignant les propositions qui ont été effectuées dans les travaux de recherche avec leurs avantages et leurs inconvénients.

1.4.1 Manque de données

Dans le cadre d'une approche de recommandation fondée sur le FC (basé sur la mémoire), l'identification des appréciations des utilisateurs est l'un des piliers de base du processus de recommandation. Elle permet en effet de modéliser les utilisateurs dans le but de prédire les futurs goûts d'un utilisateur actif en se basant sur les appréciations connues d'un groupe d'utilisateurs.

Ces appréciations sont soit renseignées explicitement par les utilisateurs eux-mêmes ou bien induites par le système sur la base de l'analyse des interactions de ces utilisateurs avec le système.
Or, dans les deux cas, souvent les données relatives aux appréciations des utilisateurs manquent et s'avèrent insuffisantes pour le bon fonctionnement du système de recommandation [Sarwar *et al.*, 2000b]. En effet, la quantité de données ou de notes disponible demeure toujours insuffisante pour pouvoir prédire correctement les notes manquantes.
Par conséquent, en raison de ce manque de données, la modélisation des utilisateurs devient complexe. Les modèles utilisateurs deviennent ainsi peu fiables, parce qu'ils ont été construits en se basant sur un volume limité de données.
En outre, dans le cadre du FC basé sur la mémoire, quand la matrice "Utilisateur x Item" est très creuse[11], le système est incapable d'identifier un nombre significatif de voisins en

[11]Par exemple, sur la base de Movielens, environ 94% de la matrice de notes est vide

s'appuyant sur les items co-notés, ce qui se répércute sur la qualité des recommandations proposées à l'utilisateur actif et sur la performance de la totalité du système.

Pour les approches de recommandation basées sur le WUM, elles ont l'avantage de ne pas requérir de notes puisqu'elles exploitent les données d'usage. Toutefois, ces approches font face également au problème de manque de données. En effet, une masse importante de données (traces d'usage) est nécessaire afin de pouvoir découvrir des motifs fiables et prédire efficacement les futurs comportements de navigation.

Plusieurs travaux de recherche ont étudié le problème de manque de données dans le cadre du FC, en examinant l'intérêt d'exploiter certaines techniques telles que les méthodes basées sur le contenu et les méthodes basées sur un modèle (le clustering, SVD, etc.) afin de remédier à ce problème.

Comme nous l'avions décrit précédemment (cf. sections 1.3.1 et 1.3.5), devant l'indisponibilité des données de notes, le contenu peut être exploité en vue d'évaluer les similarités entre items et effectuer les recommandations. Dans ce contexte, la technique basée sur le contenu peut être hybridée avec le FC basé sur la mémoire [Balabanović et Shoham, 1997] [Pazzani, 1999] [Melville *et al.*, 2002] et/ou avec une technique basée sur un modèle telle que naive Bayes [Xiaoyuan *et al.*, 2007].

Dans le cadre des méthodes basées sur un modèle et pour remédier au manque de données, le clustering (cf. section 1.3.3) a été largement utilisé. Parmi ces travaux, le clustering a été appliqué soit aux utilisateurs ou bien aux items ou bien aux deux, afin de générer des clusters d'utilisateurs ou d'items similaires dans le but de prédire les notes manquantes. Différents algorithmes de clustering ont été utilisés dans cette optique, notamment : k-means [Ungar et Foster, 1998] [Xue *et al.*, 2005], Gibbs Sampling [Breese *et al.*, 1998], ROCK [Conner et Herlocker, 1999], etc.

En outre, les techniques de réduction de dimensionnalité ont été intégrées également au processus de FC pour faire face au problème de manque de données [Sarwar *et al.*, 2000b] [Zhang *et al.*, 2005] [Gong *et al.*, 2009]. La réduction de dimensionnalité vise à représenter les données dans un espace ayant une dimension plus réduite que celle du départ.
SVD constitue une technique de réduction de dimensionnalité, qui consiste en la factorisation d'une matrice "Utilisateur x Item" [Sarwar *et al.*, 2000b] [Goldberg *et al.*, 2001] [Zhang *et al.*, 2005] [Gong *et al.*, 2009].
Outre son apport face au manque de données, l'utilisation de la technique SVD par les sysèmes de recommandation permet d'une part de produire une représentation de plus faible dimension de l'espace original "Utilisateur x Item" et de calculer les similarités utilisateur-utilisateur au niveau de cet espace réduit. D'autre part, elle met en évidence les relations latentes entre utilisateurs et items permettant le calcul des notes manquantes [Sarwar *et al.*, 2000b].
Toutefois, la complexité de SVD en temps et en mémoire est très importante, ce qui rend l'apprentissage coûteux et inapproprié pour de grandes matrices.

Les techniques citées ci-dessus parviennent à traiter le problème de manque de données, toutefois elles présentent quelques limites.
La technique basée sur le contenu permet de calculer les recommandations concernant des items peu notés dans le système. Cependant, cette technique engendre un manque de diversité du contenu des recommandations.
En ce qui concerne les techniques de clustering et de SVD, elles permettent notamment de condenser l'espace de représentation de données en supprimant les utilisateurs ou les items non représentatifs. Cependant, le risque lié à cette suppression est la perte d'information potentielle (concernant par exemple des voisins fiables), susceptible d'entraîner une dégradation de la performance du système de recommandation.

1.4.2 Démarrage à froid

Le problème de démarrage à froid se traduit par la difficulté de générer des recommandations concernant de nouveaux items ou de nouveaux utilisateurs qui viennent d'être introduits au système de recommandation. Défini comme le problème de "systemic bootstrapping" par [Rashid *et al.*, 2008], le démarrage à froid peut concerner tous les types données (concernant les utilisateurs et les items). Ce problème se produit lorsqu'il s'agit par exemple d'un nouveau service créé et pour lequel aucune donnée n'est encore disponible [Schein *et al.*, 2002]. Ainsi, le nouveau système de recommandation en question ne peut recommander aucun item, à aucun utilisateur.

Nouveaux utilisateurs

Proposer des recommandations à un nouvel utilisateur, récemment introduit au système, constitue un enjeu pour les systèmes de recommandations. Dans le cadre du FC, tant que le système n'a aucune connaissance sur les appréciations de ce nouvel utilisateur, sa modélisation reste complexe et le système de recommandation ne sera pas capable de lui proposer des recommandations personnalisées.
Dans ce contexte, l'élicitation (à travers la sollicitation de notes explicites, de critiques ou d'informations démographiques) peut se présenter comme une solution. Or, cette sollicitation directe peut entraîner l'abandon de l'utilisateur tel que décrit dans la section 1.2.
Un autre moyen d'aborder le problème de nouveauté de l'utilisateur, est de lui proposer des recommandations arbitraires dès sa première utilisation du système. Cependant, cette stratégie risque d'occasionner une insatisfaction chez l'utilisateur, au vu de la faible qualité des recommandations.
[Rashid *et al.*, 2008] présentent d'autres stratégies pour faire face au problème de nouveauté de l'utilisateur. Ces stratégies exploitent la popularité des items et l'entropie consistant à évaluer la dispersion des avis des utilisateurs sur un item.
Par ailleurs, les profils démographiques des utilisateurs (cf. section 1.2) représentent aussi un moyen de remédier au manque de données. En effet, l'information démographique peut être exploitée en vue de construire les modèles utilisateurs. Ainsi, deux utilisateurs appartenant au même segment démographique, sont considérés comme similaires [Pazzani,

1999] [Vozalis et Margaritis, 2006]. En s'appuyant sur le principe du FC, ces similarités permettent d'identifier les voisins dont les appréciations sont considérées pour le calcul des recommandations. Or, même si des utilisateurs appartiennent à un même segment démographique, ils ne partagent pas nécessairement les mêmes goûts.

Nouveaux items

Recommander de nouveaux items constitue également un enjeu de taille pour les systèmes de recommandation. Ce problème est connu sous le nom de "latence". En effet, quand un nouvel item est intégré au système, les préférences des utilisateurs par rapport à cet item ne sont pas encore disponibles. Par conséquent, le nouvel item ne sera pas impliqué dans le cadre des recommandations. Le problème de latence a, en particulier, plus d'un impact sur les systèmes qui incorporent de nouveaux items régulièrement, comme les systèmes recommandant les articles d'actualité [Sollenborn et Funk, 2002] [Burke, 2002].

Pour pallier ce problème de latence, une stratégie consiste à sélectionner aléatoirement les nouveaux items et à proposer à l'utilisateur actif d'y attribuer des appréciations. Cependant, tel que discuté précédemment, cette stratégie pourrait occasionner une lassitude chez l'utilisateur qui risque d'abandonner le système.
Une solution alternative consiste à exploiter la technique basée sur le contenu [Lang, 1995] [Krulwich et Burkey, 1996] [Billsus et Pazzani, 2000]. Cette technique est utilisée tant que les notes sur un item ne sont pas suffisamment disponibles. Quand un nouvel item est introduit, la technique basée sur le contenu évalue la similarité de son contenu avec les items disponibles afin de l'impliquer au processus de recommandation. Néanmoins, l'utilisation de la technique basée sur le contenu engendre un manque de diversité des recommandations, ce qui entrave la performance du système de recommandation (cf. section 1.3.1).
Une nouvelle technique de filtrage (basée sur le contenu) exploitant les ontologies, a été suggérée également comme une solution au problème de latence. Cette technique a été notamment utilisée par le système Entree (qui recommande des restaurants) [Burke, 2002] et le système Quickstep-Foxtrot (qui recommande des papiers scientifiques) [Middleton *et al.*, 2004]. Les méthodes d'apprentissage utilisent les ontologies mises en place dans le cadre de ces systèmes, afin de classifier et de catégoriser les items et générer les modèles utilisateurs. Or, la limite de cette technique est la nécessité de la construction préalable d'une ontologie relative au domaine de connaissance.

Les travaux de recherche présentés dans cette section ont proposé différentes approches dans le but de faire face au problème de démarrage à froid pour de nouveaux utilisateurs ou de nouveaux items. Ces approches exploitent par exemple l'élicitation, le FC exploitant l'information démographique, la technique basée sur le contenu ou sur les ontologies.
Malgré leurs intérêts, ces approches présentent quelques limites liées notamment au manque de diversité des recommandations ou à la détérioration de la qualité des recommandations (en raison de l'utilisation de l'information démographique par exemple).

1.4.3 Sélection de voisins fiables

Dans le cadre du processus de FC basé sur la mémoire (centré sur l'utilisateur), l'approche kNN permet de retrouver les k voisins les plus proches d'un utilisateur actif dans le but d'utiliser leurs avis pour générer des recommandations pertinentes à cet utilisateur actif. Ces k plus proches voisins sont considérés comme étant les voisins les plus informatifs. Ils ont en effet des appréciations similaires vis-à-vis de l'utilisateur actif, au vu de leurs opinions concernant des items notés ou consultés en commun antérieurement.

L'identification de ces voisins dans une approche kNN peut notamment reposer sur des stratégies telles que la détermination d'un seuil de similarité ou la détermination d'un seuil d'items co-notés (cf. section 1.3.2). Or, la détermination de ce type de seuil reste problématique. En effet, avec l'intégration de nouveaux utilisateurs et d'items, pour être plus fiable, le système de recommandation réinitialise le calcul des voisinages. Par conséquent, l'ensemble des k voisins les plus proches varie et son choix n'est jamais définitif. De ce fait, ces seuils doivent être adaptés au fur et à mesure de la réinitialisation du système, tout en évitant de fixer des valeurs extrêmes pour que le pouvoir prédictif du système ne soit pas faible et pour que le bruit ne soit pas engendré à cause de voisins peu pertinents.

La limite d'une telle approche est qu'elle demeure dépendante des items notés en commun afin d'évaluer le degré de similarité entre utilisateurs et de déterminer les plus proches voisins. En l'absence de ces items co-notés, aucune modélisation d'utilisateurs n'est possible et aucun voisinage fiable ne peut être sélectionné.
Dans cette optique, d'autres méthodes permettant d'identifier des similarités entre utilisateurs ont été proposées. Il s'agit d'exploiter par exemple les associations transitives afin d'établir des liens entre utilisateurs ou entre utilisateurs et items. [Papagelis *et al.*, 2005] et [Golbeck, 2009] exploitent le principe d'inférence afin d'explorer les associations entre utilisateurs dans l'objectif d'identifier des voisins potentiellement fiables, susceptibles d'améliorer la qualité des recommandations. Néanmoins, considérant que les systèmes de recommandation sont dynamiques et que la phase de calcul du voisinage requiert un temps de calcul important, l'application de ce type d'association devrait se baser sur des stratégies permettant de limiter par exemple le nombre d'utilisateurs concernés, afin de permettre le passage à l'échelle.

Par ailleurs, la notion de confiance a également été étudiée comme un moyen de détermination de voisins fiables dans le cadre des systèmes de recommandation. [Massa et Bhattacharjee, 2004] [O'Donovan et Smyth, 2005] [Papagelis *et al.*, 2005] [Golbeck, 2009] proposent en effet de considérer la confiance en prenant notamment en compte la capacité antécédente d'un voisin à fournir ou à contribuer à des recommandations pertinentes. Par exemple, en utilisant des mesures de confiance, le système proposé par [O'Donovan et Smyth, 2005] peut spécifier à un utilisateur actif u_a que "le système vous recommande la voiture Toyota Verso, cette recommandation vous a été générée par les utilisateurs u_c, u_d et u_e, ces utilisateurs ont déjà recommandé la Toyota Verso n fois dans le passé, et ces recommandations ont été fiables r fois".

Toutefois, une telle démarche de recommandation va à l'encontre du respect de la vie privée. En effet, pour appuyer la notion de confiance, ce système de recommandation se permet d'annoncer quel utilisateur a recommandé tel ou tel item et combien de fois son avis a été utilisé dans le passé dans le cadre des recommandations.
De plus, les systèmes de recommandation basés sur la confiance requièrent un retour d'expérience des utilisateurs vis-à-vis des items recommandés. En effet, les utilisateurs doivent exprimer directement leurs retours suite à la réception des recommandations et ce en évaluant chacun des items recommandés. Ce processus s'inscrit dans l'élicitation et pourrait provoquer un agacement chez l'utilisateur (cf. section 1.2). Autrement, des heuristiques pour évaluer l'intérêt vis-à-vis d'une recommandation, devraient être définies (par exemple la durée de consultation, la lecture d'une vidéo, la commande d'un produit, etc.).

La sélection des plus proches voisins est primordiale dans la mesure où la qualité des prédictions peut être influencée par la fiabilité des voisins. L'exploitation de l'approche kNN ou du principe de confiance permettent de sélectionner des voisins pertinents, mais requièrent respectivement la disponibilité des items co-notés ou le retour de l'utilisateur. L'utilisation des inférences pour identifier des voisins potentiellement fiables est une solution prometteuse, mais reste peu appropriée pour un grand volume de données.

1.4.4 Robustesse

La robustesse constitue un challenge pour toutes les applications en ligne. Devant la difficulté d'évaluer la confiance des utilisateurs utilisant les systèmes de recommandation, ces derniers demeurent vulnérables aux manipulations et aux données bruitées. En effet, il n'y pas de garantie que les données intégrées aux systèmes de recommandation reflètent les réelles appréciations des utilisateurs.
[O'Mahony *et al.*, 2006] distinguent deux catégories de données bruitées :

- Le bruit naturel : ce bruit relève du fait que l'expression des appréciations est souvent perçue par les utilisateurs comme un processus fastidieux, ce qui peut influencer la qualité des opinions attribuées par ces utilisateurs.

- Le bruit malicieux : ce bruit provient de l'insertion d'information biaisée de la part de certains utilisateurs malveillants. Une de leurs motivations par exemple est de promouvoir leur produit ou leur article en forçant le système de recommandation à générer des notes élevées pour ceux-ci et à en faire un "push", au détriment d'autres items (concurrents) présents dans le système. De plus, le bruit malicieux peut aussi bien consister à endommager la totalité du système.

[Lam et Riedl, 2004] évaluent l'impact des attaques et des données bruitées sur l'efficacité du système. Cette évaluation est effectuée en termes de vulnérabilité d'algorithmes

utilisés et de capacité prédictive du système de recommandation. En outre, [Lam et Riedl, 2004] proposent des mesures de détection d'attaques et étudient les propriétés des items attaqués.
Pour remédier au problème des données bruitées et garantir la stabilité du système, [O'Mahony *et al.*, 2006] définissent des méthodes permettant de détecter le bruit en exploitant une théorie de détection du signal et montrent la fiabilité de ces méthodes pour la garantie de la robustesse du système contre différentes stratégies d'attaques. [Mehta *et al.*, 2007] étudient l'intérêt de certaines méthodes statistiques, telles que les techniques de factorisation de matrice, pour la stabilité du système de recommandation, malgré la présence de bruit. Par ailleurs, [O'Donovan et Smyth, 2005] montrent l'importance des modèles de confiance pour améliorer la robustesse des systèmes de recommandation. L'utilisation des modèles de confiance ont ainsi un double avantage. Ils contribuent d'une part à la sélection de voisins fiables, ils permettent d'autre part de garantir la stabilité des systèmes de recommandation.

1.4.5 Précision des recommandations

L'évaluation des systèmes de recommandation constitue une étape clé dans un processus de recommandation dans la mesure où elle reflète la performance de l'intégralité du système. Pour tout système de recommandation, prédire efficacement les futures appréciations contribue à la satisfaction des besoins des utilisateurs et à leur fidélisation.

L'évaluation des systèmes de recommandation peut prendre en compte différents critères, à savoir : la précision, la couverture, la satisfaction de l'utilisateur, la robustesse, le temps de calcul, la nouveauté et la diversité des recommandations, etc. [Anand et Mobasher, 2005].
La plupart des travaux de recherche portant sur les systèmes de recommandation, évaluent la performance de leurs algorithmes en s'appuyant notamment sur le critère de précision des prédictions. La précision permet en effet d'évaluer la capacité du système à recommander des items que l'utilisateur apprécie réellement.
A travers les algorithmes proposés par les travaux cités dans ce chapitre, l'amélioration de la précision était souvent un enjeu majeur. La performance de ces algorithmes était mesurée en effet selon le degré de précision des recommandations comparée à des techniques de recommandation standards.

Il est à signaler que la qualité et la précision des recommandations est étroitement liée à la disponibilité des données sur les appréciations. En effet, quand ces données sont rares, le système ne peut générer des prédictions précises. En outre, cette qualité de recommandation dépend également de la fiabilité de l'algorithme utilisé pour l'apprentissage des modèles utilisateurs.

Les mesures utilisées pour évaluer la précision des systèmes de recommandation vont

être détaillées dans le chapitre suivant.

Conclusion

Dans ce chapitre, nous avons décrit la typologie des données susceptibles d'être exploitées dans le cadre des systèmes de recommandation. De plus, nous avons présenté un état de l'art relatif aux principales techniques de recommandation, à savoir : la technique basée sur le contenu, le FC basé sur la mémoire ou sur un modèle, les techniques du WUM ainsi que les techniques hybrides. A travers les travaux de recherche cités notamment dans ce chapitre, il s'avère que chaque technique a des apports mais également des limites. Le choix d'une technique ou d'une autre est liée à la problématique traitée, au contexte applicatif ainsi qu'à la disponibilité des données à l'entrée du système.
Après avoir présenté les principales questions de recherche auxquelles nous nous intéressons, nous allons décrire dans le chapitre suivant l'approche générique que nous proposons, notre contexte applicatif ainsi que la méthodologie expérimentale utilisée en vue d'évaluer la fiabilité des modèles de recommandation proposés à travers cette thèse.

Chapitre 2

Schéma générique, contexte applicatif et méthodologie expérimentale

2.1 Schéma générique de la recommandation

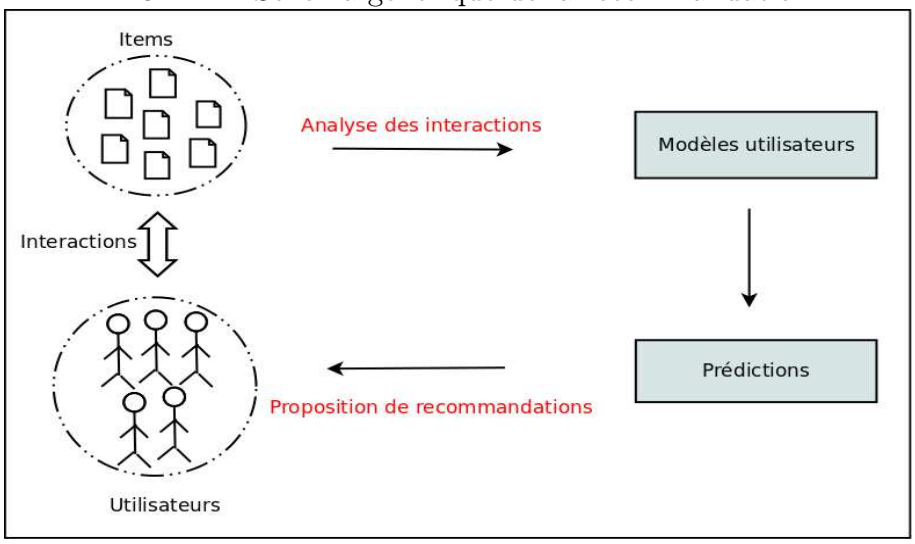
Rappelons que notre travail de recherche consiste à proposer de nouvelles approches de recommandation s'inscrivant dans le cadre d'un processus de personnalisation sur le Web. L'objectif de ces approches est d'améliorer l'accès des utilisateurs aux items au niveau des systèmes de recherche d'information, tels que les portails et les Extranets documentaires d'entreprise.

La figure 2.1 décrit le schéma générique de la recommandation auquel nous nous intéressons dans le cadre de cette thèse. A partir de l'analyse des interactions entre les utilisateurs et les items, l'objectif de ce schéma consiste à construire des modèles ou des profils utilisateurs.
Ces interactions peuvent être extraites de l'ensemble des actions effectuées sur un item par un utilisateur donné, telles que :

- une consultation d'item,
- une évaluation d'item au travers de l'attribution de note ou d'appréciation,
- une action relative à la navigation à travers des items (cf. section 1.2, chapitre 1, partie 1).

La construction des modèles utilisateurs repose notamment sur l'analyse des actions de cet utilisateur quant aux items consultés auparavant, afin de générer les prédictions des futures opinions de cet utilisateur concernant des items qu'il n'a pas encore consultés. Il s'agit de connaître les besoins de l'utilisateur en exploitant ses actions antérieures dans le but de prédire ses futurs appréciations.

FIG. 2.1 – Schéma générique de la recommandation



Une fois les prédictions générées, une liste d'items jugés pertinents, triée généralement par ordre d'importance (i.e. un classement d'items selon un ordre de pertinence estimé par le système), est proposée automatiquement à l'utilisateur qui choisit d'accepter ou non de consulter les items recommandés.
Ainsi, l'enjeu de ce schéma de recommandation est d'anticiper les besoins et de garantir la fidélisation des utilisateurs à ces systèmes grâce à la satisfaction de leurs attentes.

Dans les sections qui suivent, nous présenterons d'une part le contexte d'application liée à nos travaux de recherche. D'autre part, la méthodologie d'évaluation sera décrite en présentant à la fois les corpus de données exploités, les métriques d'évaluation utilisées pour l'évaluation des approches de recommandation que nous avons proposées ainsi que le modèle de recommandation de l'état de l'art qui nous a servi comme banc d'essai ("benchmark").

2.2 Contexte applicatif

Cette thèse s'inscrit dans le cadre du projet *PERCAL* entre le *Crédit Agricole S.A*, en particulier avec le Pôle Innovation et l'équipe de recherche KIWI[12] du LORIA.
Le Crédit Agricole représente un des leaders de la banque de proximité en France qui compte plus que 7000 agences dans son réseau (regroupées en 39 caisses régionales) et plus de 20 millions de clients en intégrant Le Crédit Lyonnais (LCL) et ses filiales internationales. A l'origine, le Crédit Agricole proposait des services financiers dans le domaine de l'agriculture, ces services se sont étendus par la suite à divers acteurs économiques incluant les particuliers, les professionnels et les entreprises.

[12]http ://kiwi.loria.fr

Le Crédit Agricole S.A constitue l'organe central du Groupe Crédit Agricole. Il est chargé notamment d'assurer le développement et la coordination des stratégies métiers en proposant les produits et les services à commercialiser et en fédérant les moyens et les compétences incluant notamment le développement d'une plate-forme informatique commune. De plus, le Crédit Agricole S.A. a un rôle de gouvernance sur les technologies et l'innovation.

Le Pôle Innovation relève de la Direction IIG (Informatique Industrielle du Groupe) du Crédit Agricole S.A. Ce Pôle est chargé de l'étude, de l'expérimentation et de la définition des modalités de mise en œuvre des technologies au sein du Groupe Crédit Agricole. Ses missions consistent à :

- assurer la veille technologique en évaluant de nouveaux produits et en proposant des solutions techniques concernant les projets du Groupe Crédit Agricole,
- assister la mise en place de nouvelles technologies,
- assurer le respect des standards et des normes concernant les nouvelles technologies introduites au Groupe Crédit Agricole.

Le projet PERCAL s'inscrit dans les perspectives du Pôle Innovation du Crédit Agricole S.A. L'objectif de ce projet consiste en effet, à proposer de nouvelles techniques de recommandation permettant de personnaliser l'accès à l'information, en prenant en compte le contexte d'un portail Extranet d'entreprise.

Ce portail, dont l'extrait est présenté dans la figure 2.2, met notamment à la disposition des utilisateurs du Groupe Crédit Agricole des informations en matière de veille technologique (nouveaux produits, nouvelles normes, nouvelles solutions technologiques, etc.) [Bertrand-Pierron, 2006]. En outre, il oriente ces utilisateurs vers les différents sites du Groupe. Ce portail est potentiellement accessible par ces trois catégories d'utilisateurs :

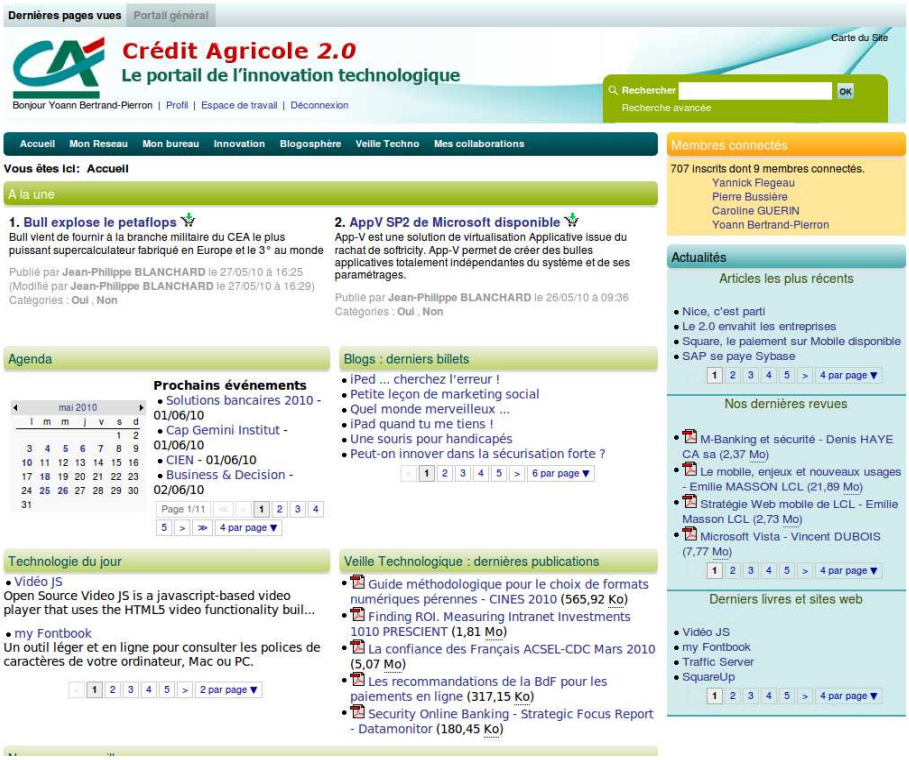
- 1100 utilisateurs qui s'authentifient actuellement sur le portail,
- 6000 à 10000 informaticiens du Groupe Crédit Agricole pouvant accéder également au portail,
- jusqu'à 150000 utilisateurs (représentant l'ensemble des employés du Groupe Crédit Agricole) pouvant consulter le portail.

Le portail Extranet est géré par l'outil "JCMS"[13] dont l'architecture fonctionnelle et technique sont présentées respectivement dans les figures 2.3 et 2.4. Basé sur Java et XML, cet outil intégre des fonctions de gestion de contenu, de gestion documentaire et de workflow, de gestion d'espaces collaboratifs et de réseau social. Suivant les rôles attribués aux utilisateurs, JCMS permet :

- la création, l'édition et la suppression de contenus,

[13]http ://www.jalios.com

FIG. 2.2 – Extrait du portail Extranet du Crédit Agricole (S.A)



- la gestion de versions de documents,
- l'indexation de contenus,
- la gestion des rôles, des droits d'accès et des circuits de validation,
- la gestion de la présentation graphique,
- la navigation et la recherche,
- le développement d'échanges et de conversations à travers des outils de réseaux sociaux.

Les items accessibles sur le portail Extranet du Crédit Agricole sont très variés, ils peuvent inclure : des articles d'actualité, des rapports techniques, des FAQ, des sondages, des blogs, des livres, etc. Leur nombre est en constante croissance. De ce fait, à partir des questions de recherche soulevées (cf. section 1.4 du chapitre précédent) et en prenant en compte ce portail Extranet, l'objectif de notre travail de recherche est de proposer de nouvelles approches de recommandation permettant d'optimiser l'usage des ressources de l'Extranet par les utilisateurs du Groupe Crédit Agricole. En effet, l'enjeu est de pouvoir mettre en place des outils de personnalisation et de recommandation collaboratifs, s'appuyant sur les usages, capables de mettre à la disposition des utilisateurs des informations pertinentes adaptées à leurs profils.

Dans le but de valider les approches de recommandation proposées à travers cette thèse, nous avons exploité des corpus de données d'usage réel et de notes explicites. Ces corpus vont être décrits dans la section suivante.

FIG. 2.3 – Architecture fonctionnelle de JCMS

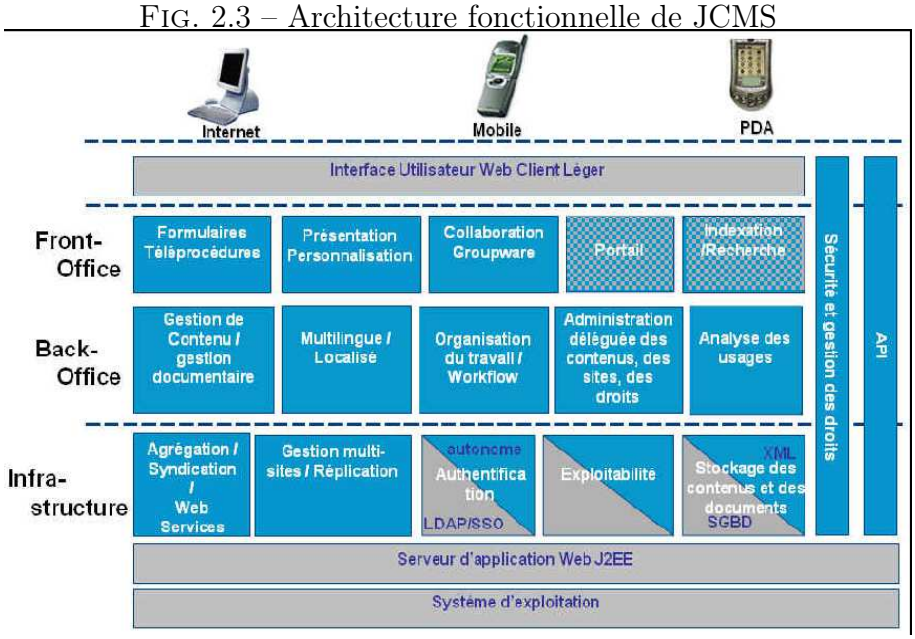
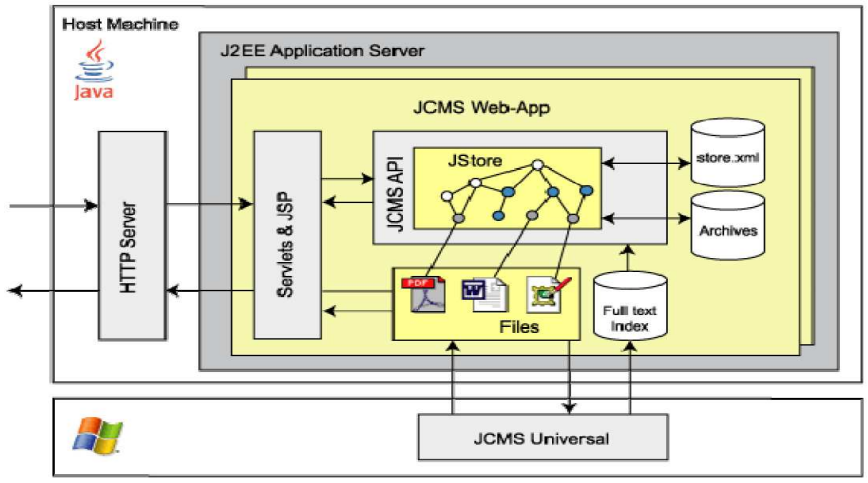


FIG. 2.4 – Architecture technique de JCMS



2.3 Données exploitées

2.3.1 Corpus d'usage

Les traces d'usage permettent de décrire l'ensemble des activités de navigation effectuées par un utilisateur sur un site Web donné.
Le WCA[14] avait publié un projet portant sur les définitions des termes relatifs aux informations contenues dans les traces d'usage. Ils concernent notamment les notions d'utili-

[14]World Wide Web Comittee web usage characterization Activity : http ://www.w3.org/wca

TAB. 2.1 – Principaux types de traces d'usage

Action	**Description**
Commander	Acheter un item
Evaluer	Noter un item
Utilisation répétée	Consulter un item d'une manière répétitive
Enregistrer/Imprimer	Enregistrer ou imprimer un item
Supprimer	Supprimer un item
Référer	Faire référence à un item
Marquer	Ajouter aux favoris
Examiner/lire	Consulter la totalité de l'item
Considérer (Temps)	Consulter le résumé
Rechercher	Rechercher un item

sateur, de page (item), de clickstream et de session.
Un utilisateur est défini comme étant un individu accédant à un fichier à partir d'un ou de plusieurs serveurs Web à travers son navigateur. Une page est représentée par tout fichier contribuant à l'affichage d'une vue sur le navigateur en un seul moment. Cette page comprend des "frames", des graphiques, des scripts, etc. Un "clickstream" est un ensemble de séries séquentielles de requêtes de pages. Une session utilisateur est représentée par les clickstreams effectués sur des pages durant une session, i.e. le moment où l'utilisateur a commencé la visite des pages Web et le moment où il a quitté le site Web en question.

Par ailleurs, [Nichols, 1997] a présenté un classement des traces d'usage dont les principaux types sont décrits dans le tableau 2.1. Ces traces concernent en particulier les sites d'e-commerce. Nous soulignons à ce niveau l'importance d'utiliser des actions telles que la commande ou la note d'un item afin d'estimer l'intérêt porté sur cet item.

D'une manière générale, à partir de la navigation, chaque activité ou demande d'affichage de page Web ou d'item de la part d'un utilisateur, génère une requête (http). Les informations relatives aux requêtes sont stockées automatiquement dans le fichier log du serveur Web. Ce fichier log constitue ainsi une importante source d'information dans la mesure où il permet de représenter le comportement navigationnel de l'utilisateur et d'inférer ses appréciations d'une manière "implicite" (contrairement à la façon explicite où l'utilisateur intervient directement pour fournir des informations sur ses appréciations vis-à-vis d'items).

Les fichiers logs peuvent être stockés sous différents formats tels que "Common Log Format (CLF)" ou "Extended CLF (ECLF)"[15]. Le format le plus courant est le CLF [Srivastava *et al.*, 2000]. Selon ce format, six informations sont stockées, à savoir :

- le nom ou l'adresse IP de la machine,
- le nom et le login HTTP de l'utilisateur,

[15]http ://www.w3.org/TR/WD-logfile.html

- la date de la requête (date, heure, écart GMT),
- la méthode utilisée dans la requête (GET, POST, etc.) et le nom ou l'identifiant de la ressource Web demandée,
- le statut de la requête,
- la taille du fichier envoyé.

Le format ECLF représente une version plus complète du CLF. Il contient en plus le nom et la version du navigateur Web, le système d'exploitation et l'adresse de la page où se positionnait l'utilisateur au moment de l'envoi de la requête.

Le nom ou l'identifiant unique de l'utilisateur n'est pas souvent une information disponible à partir d'un fichier log, surtout lorsqu'il concerne un site Web accessible aux utilisateurs sans authentification. En effet, les protocoles de communication ne peuvent pas identifier un ordinateur via l'adresse IP. En outre, les serveurs proxy peuvent regrouper plusieurs utilisateurs ou ordinateurs sous la même adresse IP. De plus, souvent ces adresses IP sont dynamiques et sont régulièrement renouvellées.

Afin de remédier à ce problème d'identification de l'utilisateur, les "cookies" peuvent être utilisés. Néanmoins, le recours aux cookies demeure problématique[16] dans la mesure où il pourrait être à l'encontre du respect de la vie privée et des données personnelles [Cooley *et al.*, 1999]. En outre, les cookies ne sont pas fiables vu que plusieurs utilisateurs peuvent utiliser un même ordinateur. Il devient ainsi complexe d'identifier un utilisateur unique.

En ce qui concerne notre contexte applicatif lié au portail Extranet du Crédit Agricole, le problème d'identification de l'utilisateur à travers les fichiers logs, n'est pas soulevé. En effet, les utilisateurs ne peuvent accéder à ce portail Extranet sans être authentifié (section 2.2).
Ces logs sont stockées sous forme de fichiers XML tel que présenté dans la figure 2.5. Ces fichiers sont générés à partir de "log4j" qui est une API de journalisation très populaire dans le monde Java.

Les principales balises contenues dans ces fichiers log sous format XML sont décrites dans le tableau 2.2. De ce fait, nous pouvons extraire à partir de ces fichiers des informations concernant notamment : l'utilisateur (balise "mid"), l'item (balise "id") et la session (balise "sessionId").

Estimation des notes

Comme nous l'avions indiqué précédemment, nous avons choisi d'exploiter l'approche par analyse des usages dans le cadre de nos modèles de recommandation. Les modèles

[16]la CNIL(http ://www.cnil.fr) met en particulier en garde contre l'utilisation des cookies, pour le profilage systématique des utilisateurs, à leur insu

FIG. 2.5 – Extrait du fichier log en format XML

```
<stat ip=127.0.0.1 startDate=1205854948962 endDate=1205854949067
port=8080 urid=ca801 method= GET referer=j6/home scheme=http
serverName=127.0.0.1 uri=/jcms/display.jsp qs=j193
sessionId=F87C360D2B3 locale=fr userAgent=Mozilla/5.0
id=j193 type=generated.PortletPortal
pub=j193 mid=j2 workspace=j4 gids=ca8015133 ccat=j5
pcat=j5 portal=j193 zone=Public name=Identification/>
```

TAB. 2.2 – Description des principales balises du fichier log du Crédit Agricole

Balise	**Description**
ip	L'IP de l'utilisateur
port	Le port de l'utilisateur
startDate / endDate	Le temps de début et de fin de consultation de l'item
method	GET ou POST du protocole HTTP
referer	L'URL source du clic
mid	Identifiant de l'utilisateur authentifié sur JCMS
id	Identifiant de l'item
type	Type de l'item consulté (Faq, News, Brèves...)
pub	Identifiant de la publication parente
name	Nom de la page
sessionId	Identifiant de la session
port	Numéro du port
locale	La langue utilisée
userAgent	Le navigateur de l'utilisateur

proposés à travers cette thèse, dont la description sera détaillée dans la deuxième et la troisième partie de ce manuscrit, sont collaboratifs, centrés sur l'utilisateur et s'inscrivent dans le cadre des approches proactives de recommandation (c.f. section 1.2).
Dans la phase de prédiction, nous avons exploité la fonction de prédiction du FC basé sur la mémoire, afin de générer des valeurs numériques de prédiction, en se basant sur les notes "implicites" des voisins identifiés *a priori*. Le choix de calculer ces valeurs relève notamment du besoin de comparer la performance de nos approches au FC basé sur la mémoire, utilisé largement par la communauté scientifique. De ce fait, pour calculer ces notes implicites, nous avons exploité les traces d'usage.

L'intérêt d'utiliser les traces d'usage pour estimer les appréciations ou les notes implicites a été déjà examiné dans quelques travaux de recherche.
[Chan, 1999] exploite en effet les traces d'usage afin d'estimer l'intérêt que porte un utilisateur sur un item ou une page Web donnée. Chan a proposé à cet effet la formule (2.1) "Page Interest Estimator" pour estimer une appréciation en prenant en compte les

indicateurs suivants :

- la fréquence ou le nombre de visites d'une page Web (Frq(Page)),
- l'ajout aux favoris d'une page Web (IsBookmark(Page)),
- la durée de consultation d'une page Web (Dur(Page)),
- la récence de visite d'une page Web (Rec(Page)),
- les liens visités sur une page Web (LinkPerc(Page)).

$$Interest(Page) = Frq(Page) * (1 + IsBookmark(Page) + Dur(Page) + Rec(Page) + LinkPerc(Page) \quad (2.1)$$

Dans le cadre de notre travail de recherche, nous nous sommes inspirés de l'étude de [Chan, 1999] pour estimer les appréciations à partir des traces contenues dans les fichiers logs du Crédit Agricole, concernant les items accessibles sur le portail Extranet. Il est à signaler qu'au départ, nous avions opté pour les indicateurs soulignés par [Chan, 1999] tels que : l'ajout aux favoris d'un item, la fréquence de consultation d'un item et la durée de consultation d'un item. Nous avions choisi en outre d'exploiter d'autres indicateurs tels que : l'envoi d'un item à un ami et l'impression d'un item. Toutefois, les informations se rapportant à certains indicateurs (comme l'ajout aux favoris, l'envoi ou l'impression d'un item) ne pouvaient pas être disponibles vu que les utilisateurs n'exploitent pas ces fonctionnalités au niveau du portail Extranet. De ce fait, nous avons retenu les indicateurs de fréquence de visite et de durée de visite d'un item.
Notons que nous n'avons pas pris en compte le critère de récence parce que nous considérons que le fait de consulter un item plus récemment qu'un autre peut être lié notamment à la date de première publication de cet item. En effet, comme la récence de visite d'un item (selon [Chan, 1999]) est évaluée notamment en fonction de la date actuelle, les items visités dont la publication est récente auront un plus grand poids au détriment des items publiés et visités précédemment. Or, ces derniers peuvent être plus pertinents pour l'utilisateur.

Le corpus de données que nous avons exploité et qui inclut ces fichiers logs, comprend 748 utilisateurs et 3856 items. Ces données ont été collectées durant les années 2007 et 2008. Depuis l'année 2008, le corpus a augmenté en incluant de nouvelles données (fichiers logs de navigation), mais pour des raisons de stabilisation d'échantillon, nous avons gardé le corpus initial.
Comme dans un processus de WUM, la première étape consiste à prétraiter les traces d'usage [Cooley *et al.*, 1999] [Han et Kamber, 2001] et à parser les fichiers logs en XML, afin d'effectuer un "nettoyage" de données (en supprimant les entrées dans les logs qui ne sont pas nécessaires à l'analyse d'usage) et de repérer :

- l'identifiant de l'utilisateur,

– l'identifiant de l'item visité,
– l'identifiant de la session,
– le temps de début et de fin de session,
– le temps de visite d'un item.

Dans une deuxième étape, pour l'estimation des notes implicites, nous avons pris en compte les indicateurs précisés ci-dessus.
La fréquence correspond au nombre de fois où l'utilisateur a consulté un item. Elle est calculée sur la base de l'équation (2.2). En considérant un utilisateur actif u_a, la fréquence de visite d'un item i_k est le ratio entre le nombre de visites de i_k ($N_{(u_a,i_k)}$) et le nombre moyen de visites de tous les items I ($\overline{N_{(u_a,I)}}$).

$$Frequency_{(u_a,i_k)} = \frac{N_{(u_a,i_k)}}{\overline{N_{(u_a,I)}}} \tag{2.2}$$

En ce qui concerne la durée, elle est calculée comme le ratio entre la durée totale de visite de i_k ($Drt_{(u_a,i_k)}$) et la durée totale de visites de tous les items I ($Drt_{(u_a,I)}$), selon l'équation (2.3). La durée de visite d'un item a été calculée à partir des informations fournies par les balises "startDate" et "endDate" contenues dans les fichiers logs. La durée maximale de visite d'un item a été fixée par un "timeout" afin d'éviter une situation où l'utilisateur ne consulte pas réellement l'item même s'il a envoyé une requête pour l'affichage de cet item.

$$Duration_{(u_a,i_k)} = \frac{Drt_{(u_a,i_k)}}{Drt_{(u_a,I)}} \tag{2.3}$$

Une fois les fréquences et les durées calculées pour chaque item, nous avons utilisé l'équation (2.4) proposée par [Castagnos, 2008] afin de pouvoir calculer et normaliser les notes selon l'échelle choisie $[1-5]$. Il s'agit de l'échelle de note la plus utilisée par les systèmes de recommandation exploitant les notes numériques.
Dans l'équation (2.4) $f_{Transf_{(u_a,i_k)}}$ désigne la fonction de transformation de la note de u_a sur i_k. v_{min} et v_{max} sont respectivement les notes minimum et maximum correspondant à l'échelle de note, i.e. 1 et 5. $p(c)$ représente le poids attribué au critère (fréquence et durée dans notre cas), $c(u_a, i_k)$ est la valeur du critère et c_{max} représente la valeur maximum du critère.

$$f_{Transf_{(u_a,i_k)}} = v_{min} + (\frac{\sum_c p(c) * c(u_a, i_k)}{\sum_c p(c)} * \frac{v_{max} - v_{min}}{c_{max}}) \tag{2.4}$$

Après la normalisation des valeurs, une matrice de notes "implicites" est générée, telles que les lignes représentent les utilisateurs et les colonnes représentent les items. Les notes

implicites obtenues suite à l'application de cette normalisation ont été validées parce qu'elles correspondent à l'échelle de notes retenue $[1-5]$ et leur répartition sur cette échelle reflète les différents degrés d'appréciation des utilisateurs concernant les items.

Pour les besoins d'évaluation, le corpus de données exploitant cette matrice a été réparti en deux corpus : un corpus d'apprentissage et un corpus de test qui comprennent respectivement 80% et 20% de données. Cette répartition a été effectuée en prenant en considération l'ordre des sessions dans les fichiers logs (i.e. pour un utilisateur donné, ses premières sessions font partie du corpus d'apprentissage alors que les sessions les plus récentes se retrouvent dans le corpus test).

Pour le corpus d'apprentissage, la matrice de note utilisée a un niveau de notes manquantes ("sparsity") de 96%. Il est calculé comme étant le rapport entre le nombre d'entrées vides et le nombre total des entrées dans la matrice (taille de la matrice) (cf. équation 2.5).

$$NiveauSparsity = \frac{NombreEntreesVides}{TailleMatrice} \tag{2.5}$$

2.3.2 Corpus de notes explicites

Dans le but d'évaluer la qualité des recommandations produites par nos approches, nous avons eu recours également au corpus de données de notes explicites "Movielens" proposé par le laboratoire de recherche Grouplens[17].

Le corpus utilisé comprend 100.000 notes attribuées par 943 utilisateurs sur 1682 films. Les valeurs de notes sont des entiers qui correspondent à l'échelle $[1-5]$. Dans ce corpus, chaque utilisateur a au moins noté 20 items.
80% de ce corpus constitue les données d'apprentissage et 20% représente les données de test. Chaque ligne du corpus représente une note d'un utilisateur sur un film en indiquant le timestamp de cette action.
Le tableau 2.3 présente des exemples de lignes de notes provenant du corpus Movielens. Dans ces lignes les informations sont présentées sous la forme suivante : utilisateur id | item id | note | timestamp.

Tab. 2.3 – Exemple de notes du corpus Movielens

Identifiant de l'utilisateur	**Identifiant de l'item**	**Note attribuée**	**Timestamp**
196	242	3	881250949
184	302	4	891717742
22	177	1	878887116

[17]http ://www.grouplens.org

Le niveau de manque de données ("sparsity") correspondant à la matrice de notes Movielens est équivalent à environ 94%.

L'utilité du corpus Movielens reste indéniable. En effet, il intègre d'une part des données de notes explicites réelles des utilisateurs, attribués à travers la plate-forme de recommandation Movielens[18] (permettant de générer des recommandations personnalisées de films). D'autre part, il est largement exploité par la communauté scientifique, d'où l'intérêt de son utilisation pour expérimenter et valider les approches de recommandation proposées en les comparant aux travaux de recherche existants.
Toutefois, l'inconvénient de l'utilisation de ce corpus est qu'il ne représente pas réellement un corpus de traces d'usage. En effet, les consultations d'items contenues dans ce corpus ne constituent pas de réelles séquences de navigation des utilisateurs. Il s'agit d'une suite d'items notés successivemennt par les utilisateurs sur la plate-forme Movielens.

2.4 Évaluation des recommandations

Afin d'évaluer la performance des systèmes de recommandation et de valider les approches de recommandation que nous proposons par rapport à des approches de l'état de l'art, différentes métriques d'évaluation sont utilisées dans la cadre des expérimentations. Le choix de telle ou telle métrique dépend notamment de la problématique de départ, des objectifs escomptés et de la nature de l'expérimentation à mener.
[Paris *et al.*, 2009] proposent une méthode d'évaluation qui prend en considération les différents acteurs dans le cadre d'une activité de recherche d'information dont notamment l'utilisateur, le système de recherche d'information et le fournisseur du contenu informationnel.
Dans le contexte des systèmes de recommandation, [Herlocker *et al.*, 2004] ont étudié les différentes stratégies d'évaluation du point de vue utilisateur, prédictions, types de corpus utilisés, etc. D'une manière générale, les différentes métriques d'évaluation évaluent la précision, la couverture, la satisfaction de l'utilisateur, la robustesse et le passage à l'échelle.

Le critère le plus évalué dans le cadre des systèmes de recommandation est la précision. La précision mesure la performance du système de recommandation en évaluant la qualité des prédictions comparées aux appréciations réelles. Les mesures de précision peuvent être soit statistiques, soit des mesures permettant l'aide à la décision.

[18] http ://www.movielens.org

2.4.1 Mesures statistiques de précision

MAE

Les mesures statistiques de précision consistent à évaluer la différence existant entre les notes prédites et les notes réellement attribuées par les utilisateurs. La mesure de précision la plus populaire pour l'évaluation des systèmes de recommandation est la MAE (Mean Absolute Error). Selon l'équation (2.6), la MAE calcule, pour chaque paire <note-prédiction>, la moyenne d'erreur absolue entre les notes prédites $Pred(u_a, i)$ et les notes réelles des utilisateurs $v(u_a, i)$. n représente le nombre d'items prédits présents dans le corpus test.
Plus la valeur de MAE est faible, plus les prédictions sont précises et le système de recommandation est performant.

$$MAE = \frac{\sum_{i=1}^{n} |v(u_a, i) - Pred(u_a, i)|}{n} \tag{2.6}$$

La MAE a été fréquemment utilisée pour l'évaluation des systèmes de recommandation et du FC [Shardanand et Maes, 1995] [Herlocker *et al.*, 1999]. L'avantage de la MAE est qu'elle est simple à utiliser, facile à interpréter et qu'elle est largement utilisée par la communauté scientifique, ce qui permet de positionner les approches de recommandation proposées par rapport aux travaux de recherche existants.
Néanmoins, pour l'évaluation de systèmes de recommandation proposant des listes ordonnées de recommandation (listes TopN), la mesure MAE peut ne pas être appropriée [McLaughlin et Herlocker, 2004].

Il existe d'autres mesures statistiques de précision évaluant les prédictions numériques, notamment : "Root Mean Squared Error", "Mean Squared Error" qui attribuent un poids plus important aux prédictions dont l'erreur est élevée, par rapport aux prédictions précises (i.e. ces deux mesures pénalisent plus que la MAE les sytèmes de recommandation générant des prédictions dont le taux de précision est faible).

HMAE

Les systèmes de recommandations ont pour objectif de calculer les prédictions des notes manquantes concernant le maximum de paires <utilisateur-item>. Une fois ces prédictions calculées, les items ne sont pas tous recommandés par la suite aux utilisateurs. En effet, seuls les items ayant les valeurs de prédiction les plus élevées sont proposées. Dans ce cas, l'erreur concernant les items ayant de faibles valeurs de prédiction n'est pas utile quant à l'évaluation de la performance des systèmes de recommandation, tandis que l'erreur relative aux items ayant des notes prédites élevées est d'une grande importance

en terme d'évaluation.
La HMAE permet en effet d'évaluer les "faux positifs" qui représentent les items jugés pertinents par le système, alors qu'ils ne le sont pas réellement (en comparaison avec le corpus test par exemple). Avec la détection des faux positifs, le système ne risque pas d'être pénalisé suite à une recommandation d'item non pertinent susceptible d'engendrer une insatisfaction chez l'utilisateur.

Afin d'évaluer la capacité d'un système de recommandation à proposer des items pertinents aux utilisateurs actifs, la HMAE (High MAE) [Baltrunas et Ricci, 2007] peut être utilisée. Selon l'équation (2.7), la HMAE est similaire à la MAE, mais elle a la particularité de considérer uniquement les prédictions élevées. Dans le cadre de nos expérimentations, nous avons pris en compte les notes $Pred'(u_a, i) \in [4-5]$ comme étant les notes élevées. m représente ici le nombre d'items prédits avec des valeurs élevées.
Plus la valeur de HMAE est faible, plus le système de recommandation est performant.

$$HMAE = \frac{\sum_{i=1}^{m} |v(u_a, i) - Pred'(u_a, i)|}{m} \tag{2.7}$$

La HMAE n'exploite pas les items ayant des valeurs de prédictions faibles, mais qui ont des valeurs réelles élevées dans le corpus test. Son avantage est sa capacité à évaluer la précision des recommandations, jugées pertinentes, qui sont effectivement suggérées aux utilisateurs.

2.4.2 Mesures permettant l'aide à la décision

Les mesures permettant l'aide à la décision consistent à évaluer jusqu'à quel point le système de recommandation peut recommander des items potentiellement pertinents pour l'utilisateur [Adomavicius et Tuzhilin, 2005] (les items susceptibles d'être très appréciés). En d'autres termes, ces mesures évaluent la pertinence des recommandations en calculant, dans une liste de recommandation, la proportion d'items qui sont effectivement utiles et pertinents pour l'utilisateur actif.

Pour les besoins d'évaluation en terme d'aide à la décision, les appréciations ou les notes des utilisateurs doivent être transformées dans le cadre d'une échelle binaire ("Aime" ou "Aime pas") afin de distinguer les items pertinents de ceux qui ne le sont pas, pour un utilisateur donné.
Ainsi, dans le cadre de nos expérimentations, un item est considéré comme pertinent lorsqu'il dispose des valeurs les plus élevées, c'est-à-dire des valeurs entre 4 et 5 sur l'échelle choisie $[1-5]$. Nous considérons que les notes de 1 à 3 correspondent à des items non pertinents pour l'utilisateur.

Les mesures permettant l'aide à la décision sont principalement issues du domaine de

la recherche d'information. Elles incluent notamment : la précision, le rappel et la mesure F1 [Herlocker *et al.*, 2004].

Précision

La précision évalue si un item sélectionné par un utilisateur est réellement perçu comme étant pertinent par ce même utilisateur [Anand et Mobasher, 2005]. Un item sélectionné représente un item qui est proposé par le système de recommandation à l'utilisateur actif et qui est contenu en même temps dans le corpus test. Le tableau 2.4 [Herlocker *et al.*, 2004] présente les catégories d'items répartis selon l'intersection entre les listes de recommandation et les appréciations réelles des utilisateurs.
A partir de ce tableau, la précision est calculée sur la base de l'équation (2.8) comme étant le rapport entre le nombre d'items pertinents sélectionnés N_{ps} et le nombre d'items sélectionnés par un utilisateur actif N_s.

$$P = \frac{N_{ps}}{N_s} \tag{2.8}$$

La précision générale du système de recommandation correspond ainsi à la moyenne des précisions calculées pour chaque utilisateur actif. Plus cette précision est élevée, plus le système de recommandation est performant.

TAB. 2.4 – Catégories d'items basées sur l'intersection entre listes de recommandation et préférences réelles

	Sélectionné (s)	**Non Sélectionné (ns)**	**Total**
Pertinent (p)	N_{ps}	N_{pns}	N_p
Non Pertinent (np)	N_{nps}	N_{npns}	N_{np}
Total	N_s	N_{ns}	N

Rappel

Le rappel mesure la probabilité qu'un item pertinent soit sélectionné par l'utilisateur actif. Il est calculé sur la base de l'équation (2.9) comme étant le ratio entre le nombre d'items pertinents sélectionnés par l'utilisateur "N_{ps}" et le nombre total d'items pertinents disponibles "N_p" [Herlocker *et al.*, 2004].

$$R = \frac{N_{ps}}{N_p} \tag{2.9}$$

Comme pour la précision, le rappel relatif à la totalité du système est évalué comme étant la moyenne des rappels calculés individuellement.

Il existe une mesure combinant la précision et le rappel [Sarwar *et al.*, 2000a]. Il s'agit de la mesure "F1". Elle représente la moyenne harmonique entre la précision et le rappel, suivant l'équation (2.10). La valeur de F1 varie de 0 à 1. Lorsque les scores de précision et de rappel sont équivalents, la qualité des recommandations est considérée comme parfaite.

$$F_1 = \frac{2PR}{P+R} \tag{2.10}$$

2.4.3 Couverture

La couverture mesure la capacité du système à fournir des recommandations. En FC basé sur la mémoire, la couverture peut être évaluée par rapport à la capacité du système de recommandation à générer des prédictions pour toutes les notes manquantes au niveau de la matrice de notes "Utilisateur x Item". Elle peut être également évaluée en prenant en considération uniquement les prédictions contenues dans le corpus test[19].
En effet, dans certains cas, le système de recommandation exploitant le FC, peut être incapable de calculer les recommandations. Cette incapacité peut notamment être engendrée par le manque de données. En effet, faute de notes provenant des voisins, le système aura des difficultés à calculer certaines prédictions.

Ainsi, un système de recommandation ne peut être performant que lorsqu'il est susceptible de calculer un nombre suffisant de prédictions concernant un maximum d'utilisateurs. Autrement dit, le système doit pouvoir répondre aux attentes des différents utilisateurs actifs présents dans le système.

2.4.4 Temps de calcul

La performance d'un système de recommandation peut être également évaluée en terme de temps de calcul. Il s'agit d'un temps de calcul réel qui permet d'évaluer le temps requis pour l'exécution des algorithmes et l'obtention des résultats escomptés.
Il va de soi que la mesure du temps de calcul est dépendante des spécifications matérielles de la machine utilisée pour l'exécution de ces calculs, ainsi que des programmes et applications lancés simultanément sur cette machine au moment des calculs.
En ce qui concerne nos expérimentations, elles ont été réalisées sur un PC DELL avec Windows Server 2003, ayant 2 Go de RAM et un processeur de 3,4 GHz (Pentium IV).

[19] L'intérêt de cette évaluation découle du fait que la qualité des recommandations est mesurée également sur le corpus test

2.5 Benchmark

La validation des approches et des algorithmes de recommandation proposés dans le cadre de cette thèse repose sur l'évaluation de la performance de ces approches comparée à des modèles de l'état de l'art. Nous avons ainsi choisi de comparer nos approches au principal modèle de l'état de l'art qui est le FC que nous allons appeler dans la suite de cette thèse "Filtrage Collaboratif Standard" (FCS) (cf. section 1.3.2).
Le FCS est une méthode de recommandation basée sur la mémoire, exploitant les données de notes afin de prédire les futures appréciations des utilisateurs.
Pour identifier les plus proches voisins, le FCS utilise le "coefficient de corrélation de Pearson" afin d'évaluer les similarités entre utilisateurs. Les voisins identifiés sont par la suite impliqués au calcul des prédictions en se basant sur la "somme pondérée".

Il est à signaler qu'au moment de la prédiction, les mêmes paramètres ont été appliqués à la fois à nos approches et au FCS afin de permettre leur comparaison. Ces paramètres concernent le seuil de similarité et le nombre d'items co-notés entre un utilisateur actif et ses voisins, permettant le choix des voisins les plus proches.

Dans ce chapitre, nous avons présenté le schéma générique de la recommandation, tel que nous le percevons. Nous avons également décrit le contexte applicatif ainsi que la méthodologie expérimentale incluant les corpus et les métriques d'évaluation utilisés pour évaluer la performance des modèles que nous avons proposés.
La partie suivante est consacrée à la présentation de l'approche collaborative comportementale de recommandation, qui représente l'une des contributions majeures de cette thèse.

Deuxième partie

Approche collaborative comportementale de recommandation

Chapitre 1

Vers un Filtrage Collaboratif Comportemental

Parmi les verrous qui entravent la performance des systèmes de recommandation, nous pouvons citer : le manque de données (de notes explicites) ainsi que la précision des recommandations (cf. section 1.4, chapitre 1, partie 1). Dans la perspective de lever ce verrou et d'améliorer la performance des systèmes de recommandation, nous avons proposé un nouveau modèle de recommandation qui repose sur un *filtrage collaboratif comportemental* centré sur l'utilisateur. Ce modèle est appellé "Behavioral Network Collaborative Filtering (BNCF)" [Esslimani *et al.*, 2008b] [Esslimani *et al.*, 2008a].
Selon la classification des approches de recommandation de [Anand et Mobasher, 2005], ce modèle s'inscrit dans le cadre des approches proactives de recommandation qui privilégient la déduction des appréciations. Ainsi, contrairement aux approches réactives, le retour de l'utilisateur et le recours à l'élicitation n'est pas nécessaire.
Ce modèle consiste à observer le comportement navigationnel de l'utilisateur et à analyser ses traces d'usage dans le but de modéliser cet utilisateur. La construction d'un modèle utilisateur dans le cadre du BNCF repose sur l'analyse du comportement afin de prédire les goûts de l'utilisateur et d'estimer l'intérêt qu'il porte à chaque item.

Le concept de comportement englobe généralement différents aspects se rapportant à l'agissement et aux réactions d'un utilisateur dans une situation donnée. Ce comportement peut être notamment représenté par les mouvements, les actions ou les expressions verbales de cet utilisateur.
Dans le contexte des systèmes d'information sur le Web (portail d'entreprise par exemple), nous entendons par comportement, l'ensemble des actions liées à la navigation de l'utilisateur à travers un site Web. Ces actions peuvent être observées à partir de (cf. section 1.2 du chapitre 1, partie 1) :

- une consultation de page Web ou d'item,
- une manipulation d'item : des actions de copier/coller, d'enregistrement ou d'im-

pression, d'ajout aux favoris, d'envoi par mail, etc.
- indicateurs externes comme l'oculométrie ("eye-tracking"),
- indicateurs de navigation : fréquence et durée de visite d'item.

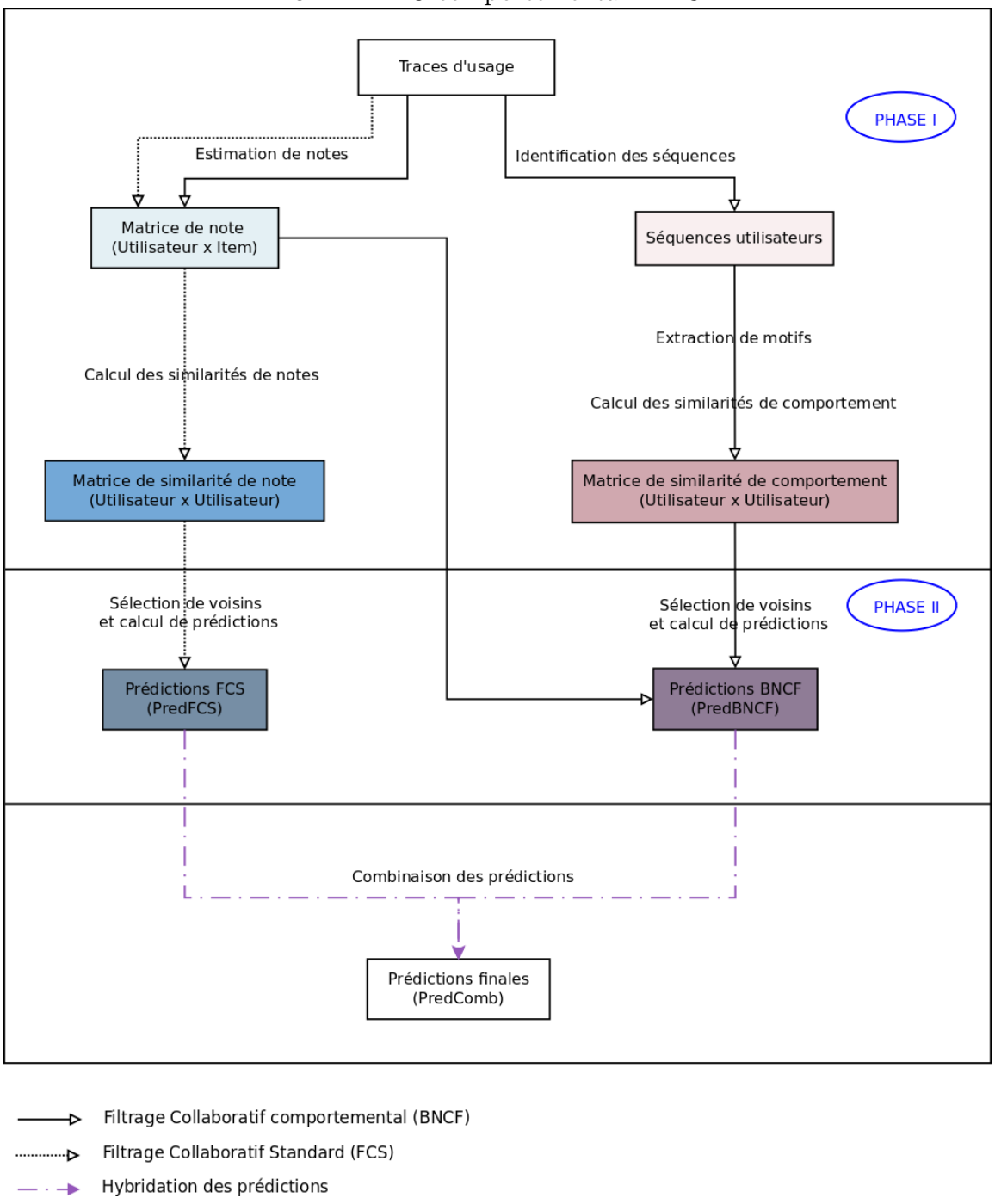
Dans le cadre du BNCF, nous considérons comme traces d'usage, les actions de consultation d'un item (en phase d'apprentissage) ainsi que les indicateurs de navigation (en phase de prédiction). Nous supposons en effet que l'analyse de ces traces est susceptible de mettre en évidence des similarités de comportement navigationnel entre utilisateurs. Le BNCF est ainsi capable d'exploiter ces similarités en vue de recommander des items à un utilisateur actif, adaptés à ses besoins.

Le modèle BNCF est inspiré à la fois des approches prédictives issues du WUM (Web Usage Mining), ainsi que des approches de recommandation basées sur la mémoire tel que le Filtrage Collaboratif Standard (FCS) [Goldberg *et al.*, 1992]. L'objectif du BNCF est de tirer profit des avantages des deux approches du WUM et du FCS, tout en remédiant aux limites qu'elles présentent.
En effet, les modèles basés sur le WUM [Anand et Mobasher, 2005] exploitent les traces d'usage dans le but d'effectuer les prédictions, grâce notamment à la découverte des motifs d'usage. Or, ces modèles requièrent une masse importante de données ou de traces afin de pouvoir extraire des motifs pertinents et de générer des prédictions fiables.
En outre, l'utilisateur n'est pas considéré pendant le processus de prédiction. Par exemple, nous supposons qu'un modèle (standard) de WUM extrait le motif fréquent noté $\langle i_8 i_5 i_2 \rangle$, signifiant que parmi toutes les traces d'usage analysées, la consultation de l'item i_2 est fréquemment produite après la séquence $\{i_8, i_5\}$ (i.e. la consultation de i_8 puis de i_5). Peu importe l'utilisateur qui aura réalisé la séquence $\{i_8, i_5\}$ pendant une session de navigation sur un site Web, l'item i_2 sera recommandé.

Quant au FCS, malgré son succès, certaines questions de recherche restent soulevées, dont notamment le manque de données explicites (les notes des utilisateurs). Par conséquent, un système de recommandation exploitant exclusivement ces données, peut être incapable de générer des prédictions adéquates s'il ne retrouve pas suffisamment de notes disponibles dans le système permettant l'identification des voisins.
A cet effet, l'exploitation du comportement dans un processus de recommandation permet d'éviter le problème de rareté des données de notes explicites. En effet, dans le cas d'une navigation sur le Web, la quantité de traces est potentiellement supérieure à la quantité de notes explicites pouvant être disponible.
En outre, l'exploitation du comportement permet de réduire le processus d'élicitation. En effet, la sollicitation directe de l'utilisateur n'est pas requise. A partir du comportement de navigation, le système est capable d'évaluer les appréciations potentielles des utilisateurs vis-à-vis d'items et même de mesurer les similarités entre utilisateurs, tel que nous le proposons ici.

La figure 1.1 décrit les différentes phases du processus de recommandation correspondant au BNCF et au FCS. Le BNCF comprend deux phases majeures : une phase

FIG. 1.1 – FC comportemental “BNCF”



d’apprentissage des modèles utilisateurs (PHASE I) inspirée du WUM et une phase de génération des prédictions (PHASE II) inspirée du FCS.
Dans la phase d’apprentissage, à partir des traces d’usage (les consultations d’items), le BNCF identifie les séquences de navigation des utilisateurs. Ces séquences sont par la suite analysées afin d’extraire les motifs d’usage. Ces motifs sont exploités en vue d’évaluer les similarités de comportement et de générer une matrice de similarité “Utilisateur x

Utilisateur". Ainsi, la particularité du modèle BNCF réside dans l'exploitation des motifs dans le but d'évaluer les similarités entre utilisateurs et non pas pour prédire directement comme dans le WUM.
La deuxième phase (PHASE II) représente la phase de prédiction. Elle vise à identifier les plus proches voisins à partir de la matrice de similarité générée et utilise leurs appréciations extraites de la matrice de notes "Utilisateur x Item" (notes estimées à partir des traces d'usage) afin de calculer les prédictions pour chaque utilisateur actif.
Le modèle FCS inclut également deux phases. La première phase (PHASE I) permet de calculer les similarités de note entre utilisateurs en exploitant la matrice de note "Utilisateur x Item". Dans la deuxième phase du FCS (PHASE II), il s'agit d'exploiter ces similarités afin de calculer les prédictions en utilisant les appréciations des voisins.
Le dernier volet de la figure 1.1 correspond à l'hybridation des prédictions générées par le BNCF et le FCS. Cette hybridation consiste à étudier l'impact de chaque modèle sur la performance du système de recommandation et à évaluer leur éventuelle complémentarité.

Ces phases vont être décrites en détails dans ce qui suit.

1.1 Extraction des motifs d'usage et calcul des similarités de comportement

La première phase du BNCF consiste en l'extraction des motifs d'usage qui vont être exploités afin de calculer les similarités de comportement navigationnel entre les paires d'utilisateurs.
Dans le cadre du BNCF, nous supposons que plus la longueur d'un motif commun à deux utilisateurs est élevée, plus ils ont un comportement similaire. Nous entendons ici par motif, une séquence fréquente, contenant une suite ordonnée d'items et qui est commune à deux utilisateurs (cf. section 1.3.4, chapitre 1, partie 1).

L'algorithme 3 présente le processus d'extraction de la longueur maximale de motifs communs. Ainsi, pour toute paire d'utilisateurs $\langle u_a, u_b \rangle$, cet algorithme exploite en entrée les séquences de navigation contenues dans leurs sessions, notées respectivement S_{u_a} et S_{u_b}, dans l'objectif d'extraire leurs motifs communs (β_i) et de calculer les longueurs correspondant à ces motifs $L(\beta_i)$. Chaque longueur correspond au nombre d'items contenus dans un motif commun. L'algorithme permet de calculer les longueurs de motifs pour chaque paire d'utilisateurs $\langle u_a, u_b \rangle$ et d'en déduire la longueur maximale des motifs L_{max} ($L_{max}(u_a, u_b) = Max(L(\beta_i))$) communs à u_a et u_b.

A la différence des motifs utilisés dans le domaine du WUM, dans notre modèle nous ne spécifions pas de support minimum déterminant un seuil pour la sélection des motifs (par exemple 20%, 30% ou 50%). En effet, l'extraction de motifs dans le cadre du BNCF est effectuée par paire d'utilisateurs, ce qui implique que le support correspondant est in-

tuitivement égal à 100%, i.e. le motif doit être nécessairement présent parmi les séquences des deux utilisateurs à la fois pour qu'il soit extrait.

Algorithm 3 Extraction de la longueur maximale de motifs

Input : Items ordonnés par sessions pour les utilisateurs u_a et u_b
Output : Longueur maximale de motifs communs
$S_{u_a} = \{ S_{1-u_a}, S_{2-u_a}, ...S_{j-u_a} \}$, $S_{u_b} = \{ S_{1-u_b}, S_{2-u_b}, ...S_{\kappa-u_b} \}$
β et γ sont deux séquences telles que : $\beta \subset S_{\mu-u_a}$ et $\gamma \subset S_{\nu-u_b}$, $\mu \in [1, ..., j]$ et $\nu \in [1, ..., \kappa]$, $\beta = \{ i_1, i_2, ...i_n \}$, $\gamma = \{ i'_1, i'_2, ...i'_n \}$
$L(\beta) = |\beta|$

for each $<u_a,u_b>$ $(u_a \neq u_b)$ **do**
    for each session de S_{u_a} **do**
        for each session de S_{u_b} **do**
            if $\exists$ entiers $ent_1<ent_2<ent_3...<ent_n$ Tel que $i_1 = i'_1$, $i_2 = i'_2$,....$i_n = i'_n$ **then**
                return $L(\beta_i)$
            end if
        end for
    end for
    return $Max(L(\beta_i))$
end for

Afin d'évaluer les similarités entre utilisateurs, nous avons proposé la nouvelle équation (1.1), permettant de calculer la similarité de navigation ou de comportement entre deux utilisateurs donnés. Cette équation prend en considération les critères suivants :

- les motifs communs entre ces deux utilisateurs,
- la longueur maximale de leurs motifs communs,
- les tailles maximales de leurs sessions.

$$SimNav(u_a, u_b) = \frac{L_{max}(u_a, u_b)}{\min(SessMax(u_a), SessMax(u_b))} \tag{1.1}$$

L'equation (1.1) calcule la similarité $SimNav(u_a, u_b)$ entre les utilisateurs u_a et u_b comme étant le ratio entre la longueur maximale de leurs motifs communs $L_{max}(u_a, u_b)$ et le minimum des tailles maximales des sessions de u_a et u_b notées respectivement $SessMax(u_a)$ et $SessMax(u_b)$. Notons que la valeur de $SimNav(u_a, u_b)$ est normalisée entre 0 et 1. Ainsi, plus la taille de L_{max} est proche des tailles de sessions des deux utilisateurs, plus $SimNav(u_a, u_b)$ tend vers 1 signifiant que u_a et u_b ont des comportements très similaires.
Il est à signaler que dans le but d'améliorer le traitement requis pour l'extraction des motifs d'usage et pour l'évaluation des similarités, nous avons évidemment réduit les paires concernées par le calcul des similarités, en considérant les relations symétriques

TAB. 1.1 – Séquences d'items de u_1 et u_2

Utilisateur u_1		**Utilisateur** u_2	
Sessions de u_1	Items	Sessions de u_2	Items
S_{1-u_1}	$i_1\ i_5\ i_{14}\ i_9$	S_{1-u_2}	$i_{12}\ i_1\ i_5\ i_8$
S_{2-u_1}	$i_2\ i_{10}$	S_{2-u_2}	$i_{20}\ i_{25}\ i_{15}$
S_{3-u_1}	$i_8\ i_{20}\ i_{13}$	S_{3-u_2}	$i_7\ i_{18}\ i_2\ i_{19}$

$(SimNav(u_a, u_b) = SimNav(u_b, u_a))$.
Au niveau de l'équation (1.1), nous avons utilisé le minimum des tailles maximales de session dans le dénominateur afin d'éviter de pénaliser un nouvel utilisateur qui a réalisé peu de sessions de faible taille et tout utilisateur qui dispose de sessions courtes. En effet, si nous considérons le maximum ou la moyenne des sessions au dénominateur au lieu du minimum, un utilisateur ayant réalisé uniquement des sessions courtes (en consultant par exemple un ou deux items par session) sera toujours faiblement similaire aux autres utilisateurs disposant de sessions de taille importante.
Cette nouvelle équation met ainsi l'accent sur l'apport des motifs d'usage pour évaluer les similarités de comportement navigationnel entre utilisateurs.

Afin d'illustrer ce processus, nous proposons l'exemple du tableau 1.1 qui présente les séquences d'items consultés par les utilisateurs u_1 et u_2 par session. En utilisant ces sessions, nous retrouvons que u_1 et u_2 ont les motifs communs suivants :

- motifs de longueur 1 ($L = 1$) : $\langle\ i_1\ \rangle\ \langle\ i_5\ \rangle\ \langle\ i_8\ \rangle\ \langle\ i_{20}\ \rangle\ \langle\ i_2\ \rangle$ (les items i_{14}, i_9, i_{10}, i_{13} étant consultés uniquement par u_1 et non pas par u_2. De même, les items i_7, i_{18}, i_{19}, i_{25}, i_{15}, i_{12} sont consultés uniquement par u_2),
- motifs de longueur 2 ($L = 2$) : $\langle\ i_1 i_5\ \rangle$,
- motifs de longueur 3 ($L = 3$) : $\emptyset$.

Ainsi, pour u_1 et u_2, la longueur maximale $L_{max}(u_1, u_2)$ de leurs motifs communs est 2 correspondant au motif $\langle\ i_1 i_5\ \rangle$ et le $\min(SessMax(u_1), SessMax(u_2))$ vaut 4. Alors, la similarité entre u_1 et u_2 est équivalente à 0.5.

En outre, en guise d'exemple de calcul des similarités par le modèle BNCF comparé au FCS entre deux utilisateurs u_3 et u_4 (notées respectivement $SimNav(u_3, u_4)$ et $SimNote(u_3, u_4)$), nous considérons le tableau 1.2 représentant les items consultés ou notés par ces deux utilisateurs. Notons que, pour des raisons de simplicité, ce tableau présente uniquement une session par utilisateur.
Nous pouvons remarquer que u_3 et u_4 ont noté en commun les items i_3 et i_5 ($I_c = \{i_3, i_5\}$), qui représentent en même temps leur motif commun le plus long ($\langle i_3 i_5 \rangle$). Ainsi, la longueur maximale correspondante est 2 ($L_{max}(u_3, u_4) = 2$).
Les tailles maximales des sessions de u_3 et u_4 (S_{1-u_3} et S_{1-u_4}) sont équivalentes respectivement à 5 et 6. Nous retiendrons 5 comme étant le minimum des tailles maximales de leurs sessions. Alors, nous pouvons calculer les similarités ainsi :

TAB. 1.2 – Items consultés par les utilisateurs u_3 et u_4

Utilisateurs	Session	Items	Notes	Moyenne de note
u_3	S_{1-u_3}	$i_1\ i_3\ i_5\ i_{10}\ i_{13}$	1 4 4 2 5	3
u_4	S_{1-u_4}	$i_3\ i_5\ i_{18}\ i_{16}\ i_{30}\ i_2$	4 4 2 2 1 5	3

– $SimNav(u_3, u_4) = \frac{2}{5} = 0.4$

– $SimNote(u_3, u_4) = \frac{\sum_{i \in I_c}(v(u_3,i)-\overline{v(u_3)})(v(u_4,i)-\overline{v(u_4)})}{\sqrt{\sum_{i \in I_c}(v(u_3,i)-\overline{v(u_3)})^2 \sum_{i \in I_c}(v(u_4,i)-\overline{v(u_4)})^2}}$
$= \frac{(4-3)(4-3)+(4-3)(4-3)}{\sqrt{((4-3)^2+(4-3)^2)((4-3)^2+(4-3)^2)}} = 1$

L'écart entre $SimNav(u_3, u_4)$ et $SimNote(u_3, u_4)$ est dû d'une part à la différence de données utilisées séparément par le BNCF et le FCS, d'autre part, à la technique permettant l'évaluation des similarités entre utilisateurs. A partir de cet exemple, nous constatons que u_3 et u_4 sont considérablement similaires en terme de notes. Or, en terme de comportement navigationnel, ces utilisateurs ne sont pas très similaires.

Cette phase (PHASE I) du BNCF, décrite dans cette section, permet de générer une matrice de similarité de comportement "Utilisateur x Utilisateur". Les voisins peuvent ainsi être identifiés et intégrés, dans une étape suivante, au calcul des prédictions.

1.2 Génération des prédictions

Une fois les similarités calculées entre utilisateurs, la deuxième phase (PHASE II) du BNCF exploite la matrice de similarité générée afin d'identifier les voisins. Les appréciations de ces voisins (récupérées à partir de la matrice de notes) sont par la suite considérées lors du calcul des prédictions.
Ces prédictions sont générées sur la base de la somme pondérée (cf. section 1.3.2, chapitre 1, partie 1), présentée dans l'équation (1.2). Cette équation est en effet la plus utilisée par les systèmes de recommandation exploitant notamment le FCS.
$SimNav(u_a, u_b)$ représente la valeur de similarité comportementale. Seuls les voisins qui sont corrélés avec l'utilisateur actif u_a (notés U_a) et ayant déjà noté l'item i_k sont considérés lors du calcul des prédictions.

$$Pred(u_a, i_k) = \overline{v(u_a)} + \frac{\sum_{u_b \in U_a} SimNav(u_a, u_b) * (v(u_b, i_k) - \overline{v(u_b)})}{\sum_{u_b \in U_a} SimNav(u_a, u_b)} \quad (1.2)$$

Dans le but d'évaluer la performance de notre nouveau modèle BNCF comparé au FCS, d'étudier leur éventuelle complémentarité et d'examiner la capacité du BNCF à

améliorer la précision des recommandations, nous avons également choisi de combiner les prédictions provenant de chacun de ces deux modèles (BNCF et FCS) dans le cadre d'un système de recommandation hybride.
Comme nous l'avons décrit dans l'état de l'art, les systèmes de recommandation hybrides combinent deux ou plusieurs techniques afin de combler les faiblesses d'une technique par une autre.
Le dernier volet de la figure 1.1 représente l'étape d'hybridation des prédictions. Ces prédictions sont combinées sur la base de l'équation (1.3) selon une méthode pondérée d'hybridation [Burke, 2002]. $PredComb(u_a, i_k)$ représente la prédiction combinée à partir des prédictions du BNCF et du FCS notées respectivement $PredBNCF(u_a, i_k)$ et $PredFCS(u_a, i_k)$, pour un utilisateur actif u_a concernant un item i_k. $\alpha \in [0-1]$ désigne le paramètre de combinaison linéaire des prédictions. Il représente le poids de chaque modèle.

$$PredComb(u_a, i_k) = \alpha * PredBNCF(u_a, i_k) + (1 - \alpha) * PredFCS(u_a, i_k) \tag{1.3}$$

Il est à rappeler que pour le calcul des $PredFCS(u_a, i_k)$, la similarité de note entre utilisateurs est utilisée. Cette similarité est calculée avec le coefficient de Pearson.
De ce fait, la principale divergence entre les modèles BNCF et FCS réside dans la phase d'apprentissage permettant le calcul des similarités.

1.3 Evaluation de la qualité des prédictions

En vue d'évaluer la qualité des prédictions générées par le BNCF et le FCS, nous avons utilisé le corpus Movielens ainsi que le corpus du Crédit Agricole décrits dans la section 2.3.2 du chapitre précédent.

Le corpus Movielens comprend 100.000 notes explicites attribués par 943 utilisateurs sur 1682 items (films). Ce corpus ne contient pas de données réelles d'usage et la notion de session n'est pas vraiment explicite. Il s'agit d'une suite d'items, qui ont été notés par les utilisateurs du système Movielens, selon des dates données ("timestamp").
A cet effet, pour l'adapter à nos besoins, nous avons considéré qu'une session correspond, dans ce corpus, à une valeur spécifique de timestamp. Or, la limite de la considération de ces timestamp est que les sessions correspondantes sont parfois très courtes. Il est à signaler que pour obtenir des motifs fiables et reflétant mieux la similarité du comportement entre utilisateurs dans le cadre du BNCF, nous n'avons pas considéré les sessions de taille 1 (i.e. des sessions où l'utilisateur a noté un seul item).
Concernant l'ordre séquentiel des items, nous avons considéré l'ordre des items tels qu'ils figurent dans le fichier qui comprend le corpus d'apprentissage.

Le corpus du Crédit Agricole constitue un corpus d'usage réel incluant les fichiers logs qui correspondent aux activités de navigation de 748 utilisateurs pouvant consulter plus de 3000 items sur le portail Extranet.
Nous avons extrait principalement les informations se rapportant aux identifiants d'utilisateurs (il s'agit d'identifiants anonymes), les séquences d'items consultés, les identifiants de sessions et le temps de début et de fin de session (cf. section 2.3.1, chapitre 2, partie 1).

Pour évaluer la précision des prédictions générées (PHASE II), nous avons utilisé les métriques d'évaluation MAE et HMAE (cf. section 2.4 du chapitre précédent).
En outre, l'objectif de cette expérimentation consiste à évaluer également la robustesse et la stabilité de notre système de recommandation en présence de notes non valides. Ainsi, dans l'expérimentation, nous avons modifié les entrées en inversant en particulier les notes ayant les valeurs de 4 et 5 dans le corpus test. Le but est d'analyser la stabilité du BNCF et du FCS en calculant la HMAE sur ce nouveau corpus. Il s'agit en effet d'évaluer jusqu'à quel point le système de recommandation peut être stable au niveau des prédictions générées et en particulier pour celles qui sont équivalentes à 4 et 5, correspondant aux items qui seront recommandés à l'utilisateur actif.

1.3.1 Résultats

Dans l'objectif d'analyser la performance de notre modèle, nous avons d'abord comparé la précision des prédictions générées par le système de recommandation en considérant que α vaut soit 0 ou bien 1. Il s'agit d'une évaluation séparée de chacun des modèles BNCF et FCS, au niveau de la qualité des prédictions en termes de MAE, de HMAE et de robustesse.

Pour la sélection des plus proches voisins au niveau des deux modèles étudiés, sur les deux corpus expérimentés, nous avons appliqué les stratégies suivantes (cf. section 1.3.2 du chapitre 1, partie 1) :

- Un seuil de similarité minimum entre un utilisateur actif et les voisins, noté θ.

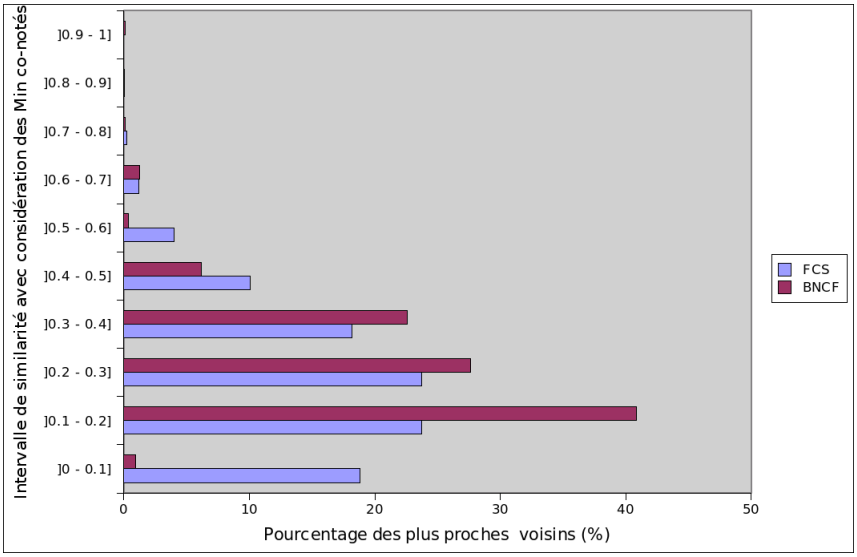
- La définition d'un nombre minimum d'items co-notés (pour le FCS) ou co-visités (pour le BNCF) entre un utilisateur actif et ses voisins. Pour toutes les expérimentations, nous avons testé d'abord différentes valeurs du nombre d'items co-notés/co-visités. Nous avons déduit que 20 permet de réaliser les meilleurs résultats de MAE (stratégie confirmée en effet par [Viappiani *et al.*, 2006]). De ce fait, nous avons retenu ce nombre en tant que paramètre de sélection des plus proches voisins.

Ainsi, tous les utilisateurs ayant co-noté ou co-visité un minimum de 20 items avec l'utilisateur actif et ayant une similarité supérieure à θ avec cet utilisateur, sont considérés

comme les plus proches voisins de cet utilisateur actif.

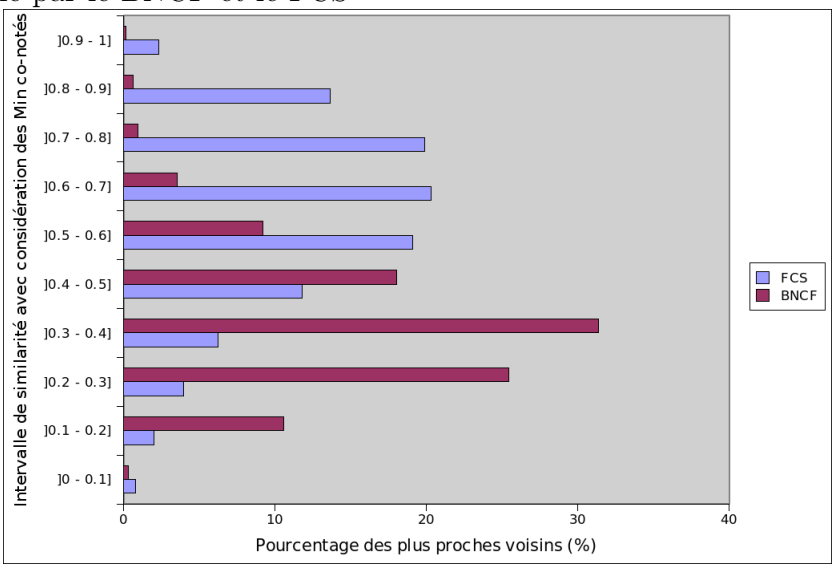
La figure 1.2 présente les pourcentages de voisins répartis selon les intervalles de similarités calculés sur le corpus Movielens, par les modèles BNCF et FCS. Les pourcentages présentés dans la figure 1.2 ont été calculés par rapport au nombre total de voisins obtenu après l'application des deux stratégies présentées ci-dessus.
En observant cette distribution, nous remarquons que la plus grande proportion de voisins calculés par le BNCF sur ce corpus (environ 90% du nombre total de voisins) ont des similarités entre 0.2 et 0.4. Pour le FCS, la répartition de ces voisins est relativement similaire à celle du BNCF, la plupart des voisins (environ 84% du nombre total des plus proches voisins identifiés par le FCS) ont des similarités entre 0 et 0.4. Notons que lorsque la similarité surpasse le seuil de 0.4 et tend vers 1, le nombre de voisins devient très faible voire nul au niveau des deux modèles.

FIG. 1.2 – Distribution des pourcentages des plus proches voisins identifiés sur le corpus Movielens par le BNCF et le FCS



La figure 1.3 présente la distribution de voisins calculés avec les modèles BNCF et FCS sur le corpus du Crédit Agricole. Au niveau de la distribution des valeurs de similarités dans la figure 1.3, nous remarquons que, comparé au FCS, le BNCF a en moyenne des valeurs de similarité plus faibles. De plus, il a également un écart type plus petit. En effet, dans le cas du FCS, il est plus facile d'obtenir de grandes valeurs de similarités si les deux utilisateurs ont des notes similaires sur 20 items co-notés. Or, dans le cas du BNCF une grande valeur de similarité suppose que les motifs communs à deux utilisateurs ont une longueur proche du minimum des tailles maximales des sessions réalisées par ces deux utilisateurs (cf. équation (1.1)).
Notons que sur le corpus Movielens (cf. figure 1.2), les voisins identifiés par le FCS ont des valeurs de similarité plus faibles que sur le corpus du Crédit Agricole. Il semblerait en effet que sur Movielens, très peu de voisins ont des notes similaires sur 20 items co-notés avec les autres utilisateurs.

FIG. 1.3 – Distribution des pourcentages des plus proches voisins identifiés sur le corpus Crédit Agricole par le BNCF et le FCS



En prenant en considération ces distributions, nous avons fait le choix d'évaluer les modèles BNCF et FCS sur les corpus Movielens et Crédit Agricole selon différents seuils θ variant de 0 à 0.4. En effet, quand θ dépasse la valeur de 0.4, le système ne peut pas retrouver suffisamment de voisins pour le BNCF et le FCS sur le corpus Movielens (cf. figure 1.2). De même, dans la figure 1.3, si nous fixons le seuil θ à une valeur supérieure à 0.4 sur le corpus du Crédit Agricole, le système va négliger une grande proportion de voisins pour le BNCF, ce qui risque de dégrader la précision des prédictions et le pouvoir prédictif du BNCF. Sur le même corpus, le FCS parvient à avoir des voisins au delà du seuil 0.4. Cependant, nous avons constaté que plus ce seuil augmente plus la couverture est faible. Ainsi, les seuils ont été choisis (de 0 à 0.4) dans le but d'évaluer la performance des modèles BNCF et FCS sur un nombre significatif de prédictions.

Résultats du BNCF et du FCS (sans hybridation)

MAE

Nous avons utilisé la MAE afin d'évaluer l'écart entre les prédictions et les notes réelles (contenues dans le corpus test). Le tableau 1.3 présente les résultats de la MAE concernant les prédictions calculées séparément par le BNCF et le FCS sur le corpus Movielens, selon la valeur du paramètre θ qui a été appliqué pour le choix des plus proches voisins.

Nous observons d'abord que les deux modèles BNCF et FCS évoluent pareillement au fur et à mesure que la valeur du seuil θ augmente. En outre, à partir des résultats du tableau 1.3, si nous considérons les résultats en MAE lorsque le seuil θ est fixé à 0.1, nous constatons que le BNCF génère des prédictions moins précises d'environ 2%, comparé aux

prédictions du FCS.

Notons que lorsque $\alpha = 1$ (en cas du BNCF) et $\theta = 0$, nous obtenons la même MAE comparé à $\theta = 0.1$, parce que nous disposons pratiquement des mêmes plus proches voisins à la base. En effet, très peu de voisins ont des valeurs de similarités inférieures à 0.1 pour le BNCF (cf. figure 1.2).
Concernant le FCS, approximativement une même précision est atteinte lorsque les valeurs du seuil θ sont situés entre 0 et 0.3. Dans ce cas, ce n'est pas nécessairement dû aux mêmes voisins qui sont impliqués. En effet, si le seuil θ augmente pour le FCS, un nombre non négligeable de voisins n'est pas pris en compte, lors du calcul des prédictions. Ainsi, le résultat similaire en MAE pour le FCS peut être expliqué par le fait que le poids associé à ces voisins n'a pas eu beaucoup d'impact sur les prédictions.

Tab. 1.3 – MAE selon la valeur du paramètre θ : corpus Movielens

Seuil θ	**FC Comportemental (BNCF)** $\alpha = 1$	**FC Standard (FCS)** $\alpha = 0$
0	**0.757**	0.741
0.1	**0.757**	**0.740**
0.2	0.760	**0.740**
0.3	0.776	0.744
0.4	0.802	0.763

Le tableau 1.4 présente les résultats de précision en MAE selon la valeur du seuil θ pour les modèles BNCF et FCS, en utilisant le corpus du Crédit Agricole.
Nous remarquons que la meilleure précision en MAE, pour les modèles BNCF et FCS, est atteinte lorsque le seuil θ est fixé à 0.2. Notons que le FCS parvient à générer des prédictions plus précises d'environ 3% comparées aux prédictions calculées par le BNCF, en considérant ce même seuil.
De plus, comme pour le corpus Movielens, lorsque les seuils 0 et 0.1 sont utilisés par le BNCF et par le FCS, le résultat de la MAE reste similaire puisque les voisins impliqués au calcul des prédictions sont approximativement les mêmes. En effet, peu de voisins ont des similarités inférieures à 0.1 sur ce corpus (cf. figure 1.3).
Lorsque le seuil est équivalent à 0.4, la précision en MAE a tendance à se dégrader respectivement pour les modèles FCS et BNCF.
Il est à signaler que pour le modèle FCS, suivant la distribution des voisins présentée dans la figure 1.3, certains voisins peuvent disposer de similarités au delà du seuil 0.4. Nous avons ainsi évalué la performance du FCS en prenant en compte d'autres seuils allant jusqu'à 0.9. Il s'est avéré que la précision en MAE s'est dégradée et la couverture tend à être très faible (perte d'environ 80% de la capacité prédictive du système). En outre, à cause de cette grande baisse de couverture, les résultats deviennent difficilement interprétables et peuvent ne pas être significatifs puisque peu de prédictions sont considérées lors de l'évaluation de la performance du FCS.

Si nous considérons les résultats en MAE obtenus sur les deux corpus, nous constatons que les résultats restent homogènes, notamment au niveau de la performance du

TAB. 1.4 – MAE selon la valeur du paramètre θ : corpus Crédit Agricole

Seuil θ	**FC Comportemental (BNCF)** $\alpha = 1$	**FC Standard (FCS)** $\alpha = 0$
0	0.799	0.772
0.1	0.799	0.772
0.2	**0.789**	**0.763**
0.3	0.790	0.774
0.4	0.847	0.779

FCS. Toutefois, l'utilisation du BNCF demeure avantageuse puisqu'il ne nécessite pas les données de notes en phase d'apprentissage, comme en FCS, afin d'évaluer les similarités entre utilisateurs. D'après les résultats, il semble en effet que les motifs d'usage exploités par le BNCF représentent des indicateurs aussi informatifs que les notes explicites et permettent d'éviter des cas où ces notes ne peuvent pas être fiables. Il s'agit notamment des cas où les utilisateurs peuvent ne pas avoir la même façon de noter les items (i.e. même s'il s'agit d'une même appréciation positive, certains utilisateurs attribuent des notes élevés et d'autres non (cf. section 1.2, chapitre 1, partie 1)).

HMAE

L'utilisation de la HMAE permet d'évaluer la performance du système de recommandation concernant la génération de prédictions ayant des valeurs élevées. Ces prédictions représentent en effet les items qui sont réellement recommandés à l'utilisateur actif.
Le tableau 1.5 compare les résultats du BNCF et du FCS en terme de HMAE, sur le corpus Movielens.
Les résultats montrent que le BNCF atteint sa meilleure performance lorsque θ est équivalent à 0.2. Ainsi, les meilleurs voisins pour le BNCF sont choisis à partir de ce seuil et ont une capacité à prédire correctement les items pour les utilisateurs actifs. Or, lorsque le seuil θ est fixé à 0.3 ou 0.4, le nombre de voisins impliqués est réduit (cf. figure 1.2). Il s'avère ainsi que la réduction des voisins engendre une déterioration de la précision en HMAE, pour le BNCF et le FCS. Notons que, comme pour la MAE, avec l'utilisation des seuils 0 et 0.1 pour le BNCF, nous obtenons les mêmes résultats en HMAE en raison de l'implication des mêmes voisins pour calculer les prédictions.
En outre, lorsque θ est fixé à 0, le FCS s'avère plus performant que le BNCF. Ce résultat induit que la stratégie d'augmentation du seuil θ pour la sélection de voisins pertinents, n'est pas appropriée dans le cadre du FCS.

TAB. 1.5 – HMAE selon la valeur du paramètre θ : corpus Movielens

Seuil θ	**FC Comportemental (BNCF)** $\alpha = 1$	**FC Standard (FCS)** $\alpha = 0$
0	0.443	**0.416**
0.1	0.443	0.444
0.2	**0.436**	0.461
0.3	0.500	0.512
0.4	0.626	0.555

Le tableau 1.6 présente les résultats en HMAE, selon différentes valeurs du seuil θ, obtenus sur le corpus du Crédit Agricole. Nous observons d'abord que, contrairement au corpus Movielens, le BNCF contribue à une meilleure performance en terme de HMAE, comparée à celle du FCS, quel que soit la valeur du seuil θ. Cette performance du BNCF est liée notamment à l'utilisation du corpus d'usage réel du Crédit Agricole permettant d'extraire des motifs fiables et d'identifier efficacement les voisins.
De plus, nous constatons que la précision du FCS est plus influencée par l'augmentation de θ, dans la mesure où la HMAE correspondant au FCS se détériore plus qu'en BNCF, au fur et à mesure que θ augmente.
Notons que les voisins ayant des valeurs de similarité entre 0 et 0.3 prédisent de la même façon les items pour le modèle BNCF, puisqu'une même HMAE est atteinte.
Comme pour l'évaluation en MAE, nous avons testé la performance du modèle FCS en HMAE au delà du seuil 0.4 considérant la distribution des voisins présentée dans la figure 1.3. Nous avons constaté que la HMAE se dégrade au fur et à mesure que le seuil augmente jusqu'à 0.9. Rappelons que l'application de cette stratégie de sélection de voisins se répercute également sur la couverture. En effet, sur le corpus Crédit Agricole, le système de recommandation fondé sur le FCS génère peu de prédictions lorsque θ est supérieur à 0.4. De ce fait, nous avons choisi d'effectuer l'évaluation en robustesse présentée dans la section suivante, en considérant le seuil θ entre 0 et 0.4.

Tab. 1.6 – HMAE selon la valeur du paramètre θ : corpus Crédit Agricole

Seuil θ	**FC Comportemental (BNCF)** $\alpha = 1$	**FC Standard (FCS)** $\alpha = 0$
0	**0.501**	0.545
0.1	**0.501**	0.545
0.2	**0.501**	**0.541**
0.3	0.502	0.571
0.4	0.528	0.588

Robustesse

Dans l'objectif d'évaluer la robustesse du système de recommandation, nous avons examiné la performance des modèles BNCF et FCS en terme de HMAE en utilisant le nouveau corpus de test, contenant les entrées erronées. Notons que nous avons maintenu les mêmes stratégies pour la sélection des plus proches voisins pour le BNCF et le FCS.
Les tableaux 1.7 et 1.8 présentent les résultats en HMAE selon différentes valeurs du seuil θ, en utilisant respectivement le corpus Movielens et le corpus du Crédit Agricole.

Les résultats du tableau 1.7 montrent que le BNCF est relativement robuste malgré la présence de données erronées dans le corpus. Comparé aux résultats du tableau 1.5, nous constatons que le BNCF garde la même évolution. En outre, le BNCF s'avère plus stable que le FCS, si nous comparons en particulier les résultats en cas de $\theta = 0$ dans les deux tableaux 1.5 et 1.7. En effet, la HMAE relative au BNCF augmente d'environ 18%, le FCS reste moins robuste vu que la HMAE correspondante augmente d'environ 24%.

TAB. 1.7 – Robustesse évaluée en HMAE selon la valeur du paramètre θ : corpus Movielens

Seuil θ	**FC Comportemental (BNCF) $\alpha = 1$**	**FC Standard (FCS) $\alpha = 0$**
0	0.542	**0.546**
0.1	0.542	0.547
0.2	0.544	0.551
0.3	0.542	0.554
0.4	**0.536**	0.559

TAB. 1.8 – Robustesse évaluée en HMAE selon la valeur du paramètre θ : corpus Crédit Agricole

Seuil θ	**FC Comportemental (BNCF) $\alpha = 1$**	**FC Standard (FCS) $\alpha = 0$**
0	0.498	0.454
0.1	0.498	0.454
0.2	0.498	0.458
0.3	0.499	0.428
0.4	**0.471**	**0.411**

A partir du tableau 1.8, contrairement au corpus Movielens, nous constatons que, comparé au BNCF (lorsque le seuil θ est fixé à 0.4) et comparé aux résultats en HMAE du tableau 1.6, le FCS contribue à une meilleure robustesse du système de recommandation. En outre, nous observons que l'augmentation du seuil de similarité n'a pas beaucoup d'effet sur la robustesse et la stabilité du BNCF. En effet, des valeurs similaires de HMAE ont été atteintes, en particulier lorsque θ est fixé entre 0 et 0.3.

Sur le corpus Movielens, le BNCF est moins sensible aux données bruitées, ce qui garantit la fiabilité et la qualité des prédictions et la non vulnérabilité du système de recommandation. Sur le corpus du Crédit Agricole, malgré la meilleure performance du FCS, la robustesse du BNCF reste généralement assez stable.

Nous pouvons déduire de cette expérimentation que la robustesse demeure influencée par la nature du corpus.

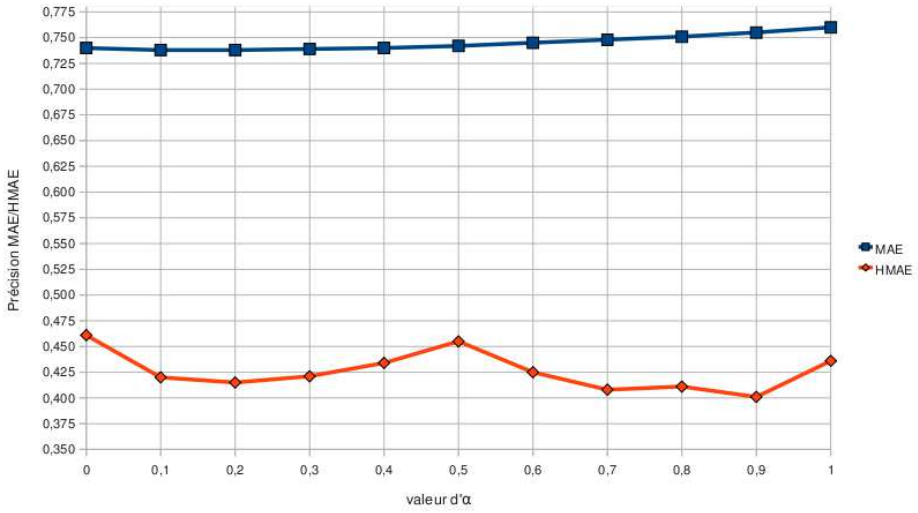
Résultats d'hybridation du BNCF et du FCS

Dans cette section, nous nous intéressons à l'évaluation de la performance du système de recommandation hybride combinant les prédictions du BNCF et du FCS. Cette évaluation a été effectuée en termes de MAE et de HMAE en utilisant les corpus de Movielens et du Crédit Agricole.
Nous avons utilisé différents poids représentés par le paramètre α. Nous avons également pris en compte les stratégies utilisées dans les tests précédents pour la sélection des plus proches voisins, en fixant le nombre minimum des items co-notés ou co-visités à 20 et le

seuil θ à 0.2 en considérant les résultats atteints avec ce seuil (cf. tableaux 1.3, 1.4, 1.5 et 1.6).

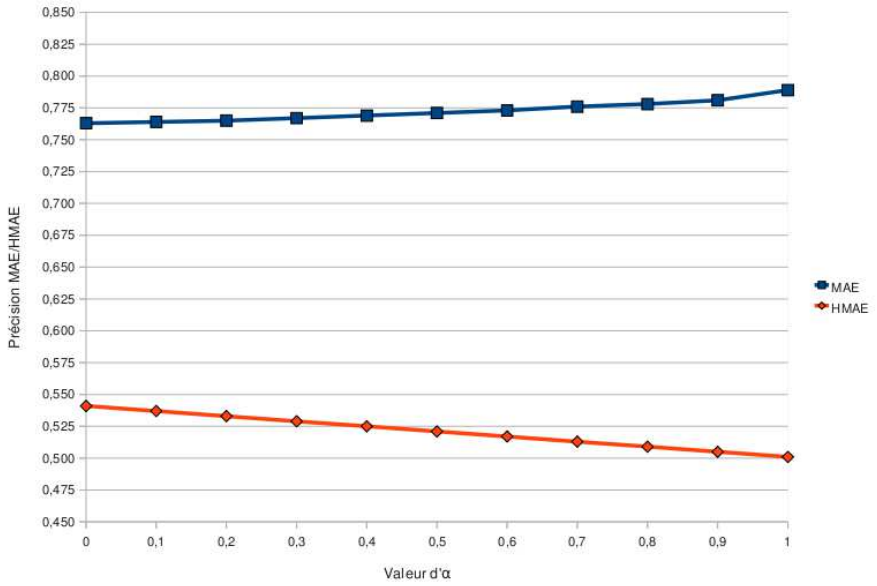
La figure 1.4 présente les résultats de cette expérimentation en termes de MAE et de HMAE sur le corpus Movielens. Nous observons d'une part que la combinaison pondérée des prédictions contribue d'une manière générale à une légère amélioration de la performance en terme de MAE. En effet, comparée aux résultats du tableau 1.3 (où 0.74 était le meilleur taux de MAE atteint), en cas d'hybridation le meilleur résultat de MAE est d'environ 0.73.
De plus, si nous comparons les résultats en MAE et en HMAE dans la figure 1.4, nous pouvons observer que la MAE atteint de meilleurs scores de précision lorsque le FCS est plus impliqué dans le calcul de la prédiction finale (par exemple lorsque $\alpha = 0.1$). Or, nous obtenons généralement la meilleure précision en HMAE, lorsque le BNCF a la pondération la plus importante (par exemple lorsque $\alpha = 0.9$). A cet effet, le BNCF reste plus adéquat pour la proposition de recommandations potentiellement pertinentes à un utilisateur actif.

FIG. 1.4 – Résultats en MAE et en HMAE sur le corpus Movielens



La figure 1.5 présente les résultats d'hybridation du BNCF et du FCS en termes de MAE et de HMAE sur le corpus du Crédit Agricole. Nous remarquons que l'évolution des résultats pour le corpus du Crédit Agricole est à peu près similaire à l'évolution des résultats en cas d'utilisation du corpus Movielens en particulier pour la MAE (cf. figure 1.4). En effet, lorsque le FCS a le poids le plus important (i.e. α tend vers 0), les prédictions calculées par le système de recommandation hybride sont plus précises en terme de MAE. Or, la meilleure HMAE est atteinte lorsque le BNCF a le poids le plus important (i.e. α tend vers 1) dans le calcul de la prédiction finale.
Ainsi, comme pour Movielens, l'importante implication du BNCF au niveau des prédictions combinées, permet de générer des recommandations appropriées aux utilisateurs actifs.

FIG. 1.5 – Résultats en MAE et en HMAE sur le corpus Crédit Agricole



En outre, nous avons réalisé un autre test afin d'évaluer la stabilité du système de recommandation hybride sur les corpus du Crédit Agricole et de Movielens, en considérant le corpus de test contenant les données erronées lors de l'évaluation.
Considérant les résultats d'hybridation du BNCF et du FCS, présentés dans les figures ci-dessus, nous avons sélectionné en particulier les prédictions combinées lorsque la valeur d'α est fixée à 0.9 suivant l'équation (1.3) (lorsque $\alpha \neq 0$ et $\alpha \neq 1$). Il s'agit en effet du meilleur résultat obtenu en terme de HMAE, sur les deux corpus, concernant le système de recommandation hybride (0.401 pour Movielens et 0.505 pour le Crédit Agricole).

Le tableau 1.9 présente les résultats d'évaluation de la robustesse du système hybride, en cas d'utilisation des deux corpus. Le résultat de la HMAE a atteint une précision de 0.494 pour le Crédit Agricole et 0.548 pour Movielens. Comparé aux résultats des tableaux 1.7 et 1.8, en particulier lorsque θ vaut 0.2, nous constatons que la robustesse demeure approximativement stable au niveau des deux corpus. Ce résultat confirme ainsi que les données erronées n'ont pas d'effet sur la robustesse du système de recommandation hybride, en particulier pour le BNCF qui dispose d'un poids important dans cette expérimentation.

TAB. 1.9 – Robustesse des prédictions combinées : corpus Crédit Agricole et Movielens

	Crédit Agricole	**Movielens**
Robustesse (HMAE)	0.494	0.548

1.3.2 Discussion

Le modèle BNCF a été proposé afin de modéliser les utilisateurs sur la base de l'analyse du comportement navigationnel. Ainsi, des utilisateurs ayant en commun des motifs d'usage, sont considérés comme similaires et partagent potentiellement les mêmes appréciations. L'exploitation des motifs d'usage dans le cadre du BNCF, permet de faire face au problème de rareté de données de notes explicites et de réduire l'élicitation.
En effet, le BNCF ne requiert pas de données de notes dans la phase d'apprentissage tel qu'en FCS. De plus, le BNCF prend en considération les traces d'usage, non pas pour prédire directement comme dans le WUM, mais pour évaluer les similarités entre utilisateurs.

Les différentes expérimentations présentées dans ce chapitre, avaient pour objectif d'évaluer l'impact du BNCF (comparé au FCS) sur la performance du système de recommandation en termes de MAE, de HMAE et de robustesse, en utilisant deux corpus différents (corpus du Crédit Agricole et de Movielens).

Si nous comparons les résultats obtenus sur les deux corpus, nous constatons que les résultats restent globalement homogènes, en particulier en termes de MAE, de robustesse et de l'hybridation des prédictions.

Au niveau du corpus du Crédit Agricole, le BNCF contribue à une meilleure précision en HMAE, en considérant l'évaluation du BNCF séparé (cf. tableau 1.6) et de l'hybridation des prédictions avec une pondération importante pour le BNCF (cf. figure 1.5).
Quant au FCS, en utilisant le même corpus, ses meilleures performances ont été obtenues en termes de MAE (cf. tableau 1.4) et de robustesse (cf. tableau 1.8).

Lorsque les modèles sont expérimentés sur le corpus Movielens, le FCS parvient à générer des prédictions précises en termes de MAE et de HMAE. Or, en exploitant ce corpus, le BNCF s'avère plus robuste et moins vulnérable face aux données bruitées, en considérant l'évolution de la HMAE correspondant au BNCF et au FCS présentée dans les tableaux 1.5 et 1.7.

Il s'avère ainsi que le BNCF demeure globalement plus performant en cas d'hybridation des prédictions (avec une importante pondération pour le BNCF) pour les deux corpus et en terme de HMAE en cas d'utilisation du corpus d'usage. En effet, ce corpus d'usage permet au BNCF d'identifier des motifs fiables permettant de retrouver des voisins pertinents contribuant à une meilleure précision en HMAE.
La robustesse et la stabilité du système de recommandation exploitant le BNCF ou le FCS, est très influencée par la nature du corpus utilisé.

Nous pouvons déduire des résultats de ces expérimentations que les traces d'usage sont une source d'information fiable permettant au système de recommandation de modéliser efficacement les utilisateurs et de générer des prédictions potentiellement pertinentes.

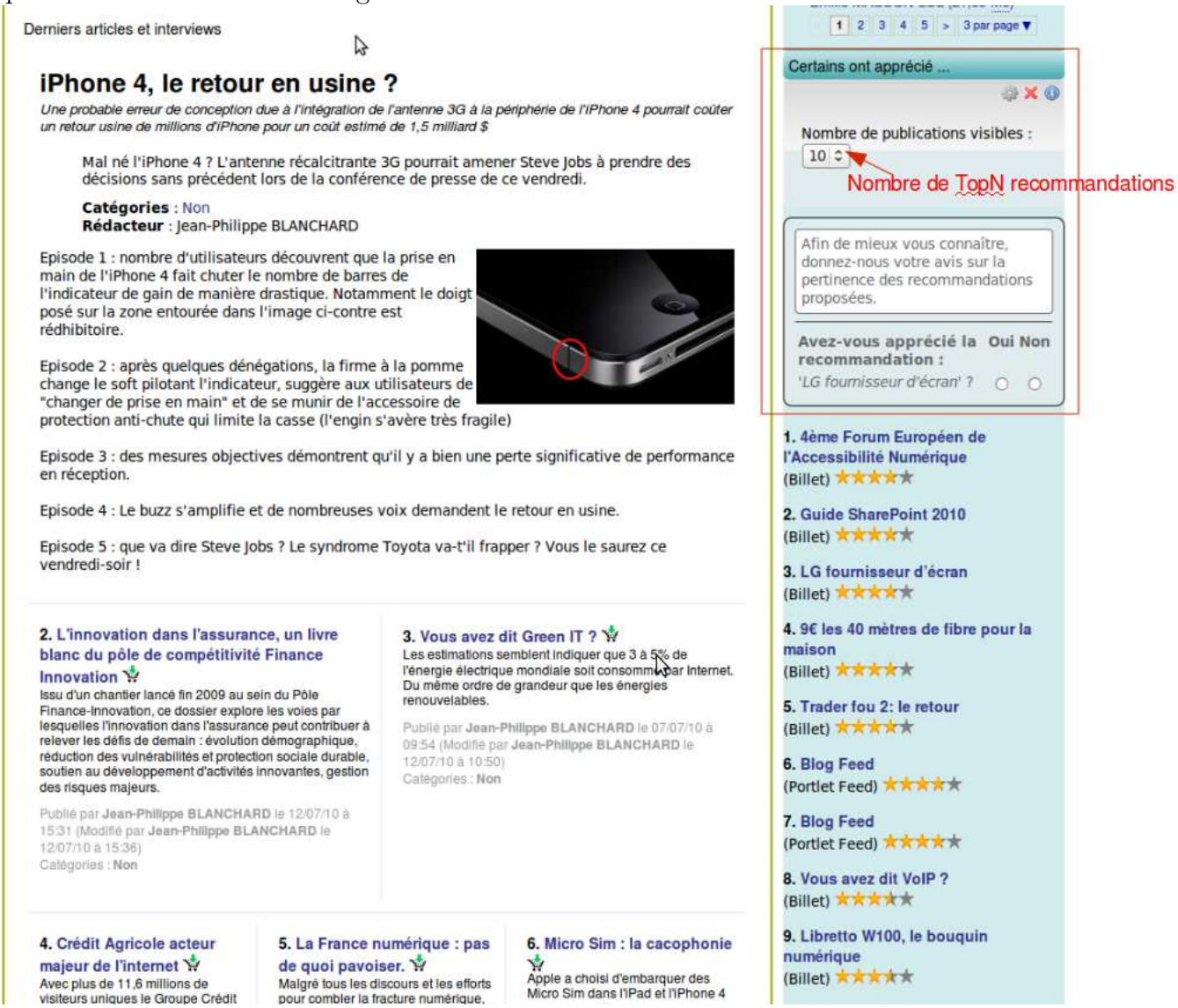
Ainsi, il serait judicieux dans les prochaines expérimentations d'évaluer la performance des modèles exploitant les motifs d'usage, sur le corpus du Crédit Agricole puisqu'il intègre des traces d'usage réelles, contrairement à Movielens.
En outre, à partir des résultats présentés dans ce chapitre, le BNCF s'avère plus approprié pour la recommandation d'items dans le cas de la navigation sur le Web en s'appuyant sur l'analyse de données implicites (des usages) telle que les données d'usage de l'Extranet du Crédit Agricole. Or, pour la recommandation d'items sur des applications de type e-commerce, le modèle FCS peut être performant à condition que les données de notes explicites soient suffisamment disponibles dans le système.

Au niveau de l'hybridation des prédictions présentée dans ce chapitre, au delà de son apport pour l'évaluation de l'impact des deux modèles sur la performance du système de recommandation, l'intérêt de cette hybridation serait l'amélioration du pouvoir prédictif du système de recommandation.
En effet, le BNCF et le FCS peuvent génèrer des recommandations pour différentes paires $< utilisateur, item >$, puisqu'ils utilisent au moment de la prédiction des voisinages différents. Il s'agit de voisinages calculés soit à partir des similarités de motifs d'usage ou bien à partir des similarités de notes. Bien que cette hybridation requiert des calculs plus importants et des paramètrages supplémentaires, elle a l'avantage de produire potentiellement des recommandations, en cas d'incapacité de l'un des deux modèles BNCF ou FCS à les génèrer.
Dans le but d'améliorer cette phase d'hybridation, une stratégie consisterait par exemple à automatiser le processus de combinaison des prédictions en fonction des données disponibles et du contexte d'utilisation du système de recommandation.

Il est à signaler qu'en collaboration avec la société Sailendra S.A.S[20], le modèle BNCF a été intégré au niveau de la plate forme CASA du Crédit Agricole contenant les outils applicatifs du portail Extranet du Groupe. Les figures 1.6 et 1.7 représentent des aperçus des recommandations sur ce portail.
La figure 1.6 est un aperçu du menu de personnalisation des recommandations proposé aux utilisateurs du portail Extranet. Les utilisateurs peuvent notamment paramétrer le nombre de recommandations à afficher.
La figure 1.7 est un aperçu de la liste (TopN) de recommandations proposée à un utilisateur, triée par ordre de pertinence (ordre estimé par le système). Notons que l'utilisateur a la possibilité d'exprimer son avis concernant les recommandations proposées par le système.
Le BNCF est actuellement testé au niveau du site Extranet du service de veille stratégique avant d'être déployé au niveau de tout le portail. Ainsi, après cette phase de déploiement, il serait pertinent d'avoir les retours des utilisateurs du Crédit Agricole suite aux recommandations proposées par notre système de recommandation. En effet, ces retours vont nous permettre d'évaluer directement la qualité des recommandations ainsi que le degré de satisfaction des utilisateurs.

[20]http ://www.sailendra.fr/

FIG. 1.6 – Aperçu du menu de personnalisation des recommandations par les utilisateurs du portail Extranet du Crédit Agricole



Par ailleurs, au delà du contexte applicatif, en vue de réduire l'espace de recherche des voisins, il serait judicieux d'étudier l'intérêt des méthodes de clustering, notamment pour la limitation du nombre de paires d'utilisateurs impliquées lors du calcul des similarités. Dans le chapitre suivant, il est question en effet d'examiner l'apport du clustering dans le cadre du BNCF. Ce chapitre est dédié à la description de cette contribution.

FIG. 1.7 – Aperçu des recommandations générées par le BNCF au niveau du portail Extranet du Crédit Agricole

Chapitre 2

Clustering en Filtrage Collaboratif Comportemental

Les expérimentations du chapitre précédent pour l'évaluation du modèle BNCF nous mènent à aborder les enjeux suivants : l'amélioration de la précision des recommandations et la réduction de l'espace de recherche pour l'identification de voisins dans un but de passage à l'échelle. C'est dans cette optique que nous avons proposé une nouvelle approche de recommandation nommée "BNCF-PAM Clustering on Similarities" (BNCF-PCS) [Esslimani *et al.*, 2009a]. Pour atteindre les objectifs cités ci-dessus, cette nouvelle approche exploite notamment un clustering d'utilisateurs.

Le clustering est une technique permettant de grouper des objets en clusters, tel que les objets appartenant au même cluster sont similaires. Dans le contexte des systèmes de recommandation, le clustering peut être appliqué aux utilisateurs ou bien aux items [Ungar et Foster, 1998]. L'avantage d'utiliser le clustering dans un processus de recommandation est de permettre à la fois de réduire l'espace de recherche pour l'identification des voisins et de pallier les problèmes de manque de données et de passage à l'échelle [Sarwar *et al.*, 2002], [Tang et McCalla, 2003], [Xue *et al.*, 2005], [Jiang *et al.*, 2006].

Les méthodes de clustering les plus exploitées par les systèmes de recommandation sont les méthodes de partitionnement dont k-means [MacQueen, 1967] est la plus populaire. Cette méthode a l'avantage d'être efficiente et permet le passage à l'échelle. Toutefois, la méthode k-means demeure peu robuste. Ce manque de robustesse est dû à sa sensibilité aux données aberrantes ("outliers") [Wang et Shao, 2004] (cf. section 1.3.3, chapitre 1, partie 1).
De ce fait, nous avons choisi d'exploiter la méthode de clustering PAM (Partitioning Around Medoid) qui est une méthode de type "k-medoïde" [Han et Kamber, 2001]. Habituellement, le clustering peut être exploité dans le cadre du Filtrage Collaboratif Standard

(FCS)[21] afin de générer des clusters en fonction des similarités de notes entre utilisateurs sur leurs items co-notés. Le clustering PAM (appliqué dans le cadre du BNCF-PCS) a pour particularité de générer des clusters d'utilisateurs en s'appuyant sur les similarités de voisins et non pas sur les similarités de notes. Ainsi, les utilisateurs sont regroupés en des clusters homogènes suivant le principe des voisinages communs, ce qui permet de considérer également les items non co-notés.
Notons que PAM a l'avantage d'être plus robuste que k-means dans la mesure où elle permet de réduire la sensibilité aux données aberrantes [Han et Kamber, 2001] (cf. section 1.3.3, chapitre 1, partie 1).

De plus, comme le BNCF, le BNCF-PCS exploite les traces d'usage afin de générer une matrice de notes implicites. Dans le BNCF-PCS, cette matrice est exploitée d'une part afin de sélectionner des "sous-séquences positives" pour chaque utilisateur. D'autre part, elle est utilisée lors du calcul des prédictions. Ces sous-séquences positives comprennent uniquement les items positivement appréciés par les utilisateurs. L'objectif est d'évaluer les similarités de comportement entre utilisateurs en se basant sur ces sous-séquences. En effet, nous supposons qu'un motif commun incluant des items appréciés positivement peut être révélateur d'une forte similarité entre utilisateurs. En outre, l'utilisation de ces sous-séquences positives va permettre de réduire l'espace de recherche lors de l'extraction des motifs requis pour l'évaluation des similarités.

Ainsi, la particularité du BNCF-PCS, comparé au BNCF, réside dans l'intégration d'une étape de clustering d'utilisateurs et dans la considération des séquences positives pour l'évaluation des similarités entre utilisateurs au niveau des clusters créés. Comme nous l'avons indiqué ci-dessus, l'enjeu de l'intégration de ces étapes est de réduire l'espace de recherche pour l'identification de voisins pertinents susceptibles de promouvoir la qualité des recommandations.

Le schéma décrivant le modèle BNCF-PCS est présenté dans la section suivante.

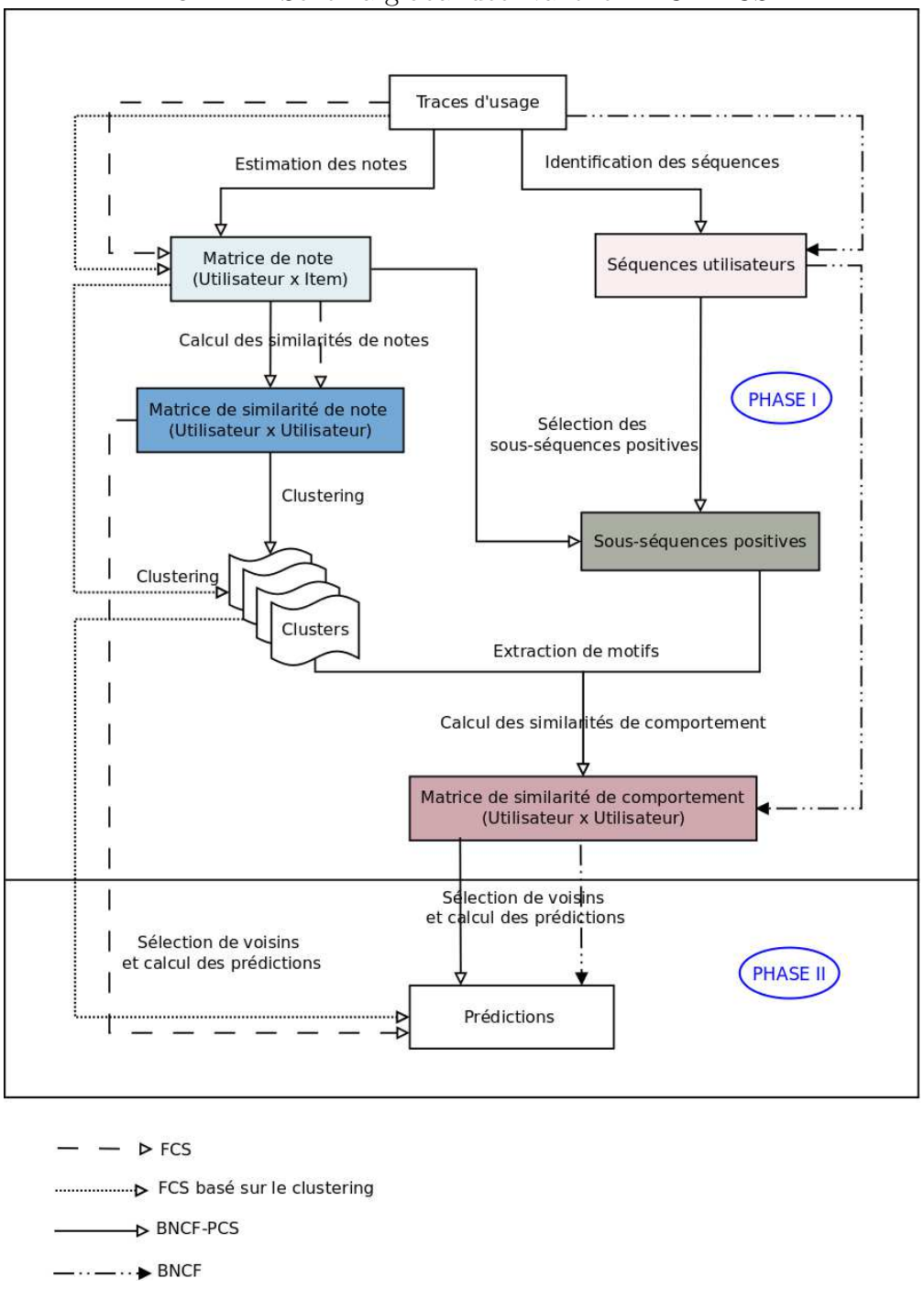
2.1 Schéma du modèle BNCF-PCS

Le BNCF-PCS est constitué des mêmes phases que celles du BNCF décrites dans la figure 1.1 du chapitre précédent. Il inclut en outre, de nouvelles étapes correspondant en particulier à la "PHASE I" qui représente la phase d'apprentissage.
La figure 2.1 décrit le schéma du modèle BNCF-PCS. Ce schéma reprend également les étapes qui s'inscrivent dans le cadre du BNCF, du FCS et du FCS basé sur le clustering (indiqué ci-dessus).

Le FCS exploite directement la matrice de note "Utilisateur x Item" contenant les notes estimées à partir des traces d'usage dans le but d'identifier les voisins et de calculer

[21] nous l'appellerons le FCS basé sur le clustering

FIG. 2.1 – Schéma global décrivant le BNCF-PCS



les prédictions. Le FCS basé sur le clustering utilise cette même matrice afin de générer des clusters d'utilisateurs selon les similarités de note et calculer les prédictions à partir de ces clusters.
Le BNCF, comme nous l'avons décrit dans le chapitre précédent, exploite les séquences de navigation (extraites des traces d'usage) en vue de calculer les similarités de comportement entre utilisateurs. Ces similarités sont par la suite exploitées pour le calcul des prédictions.

Le BNCF-PCS exploite, quant à lui, une matrice de similarité de note "Utilisateur x Utilisateur", calculée à partir des similarités de notes, dans le but de créer des clusters d'utilisateurs.
En parallèle, le BNCF-PCS sélectionne les sous-séquences positives à partir des séquences d'utilisateurs. Ces sous-séquences intègrent uniquement les items appréciés positivement par les utilisateurs. Ces appréciations sont extraites de la matrice de notes estimées à partir des traces d'usage.
Le BNCF-PCS calcule par la suite les similarités de comportement entre utilisateurs au sein de chaque cluster créé en se basant sur les sous-séquences positives.
Dans la deuxième phase du BNCF-PCS (PHASE II), les plus proches voisins sont identifiés et leurs appréciations sont combinées pour le calcul des prédictions.

Les sections qui suivent décrivent davantage le processus de recommandation dans le cadre du BNCF-PCS.

2.2 Génération des clusters

Nous avons choisi d'intégrer le clustering afin de permettre de réduire l'espace de recherche des voisins et de promouvoir la qualité des recommandations.
Le modèle FCS basé sur le clustering utilise la matrice "Utilisateur x Item" pour la génération de clusters. Ainsi, les clusters sont créés en considérant les items co-notés entre utilisateurs.
Dans le cadre du BNCF-PCS, nous avons fait le choix d'exploiter une matrice "Utilisateur x Utilisateur" (une matrice de similarité de note entre utilisateurs) pour la création de clusters. Pour la génération de cette matrice de similarité entre utilisateurs, comme dans le FCS, le coefficient de corrélation de Pearson [Herlocker *et al.*, 1999] a été utilisé afin d'évaluer les similarités de notes entre chaque paire d'utilisateurs $\langle u_a, u_b \rangle$ en se basant sur les items consultés en commun.
Notons que ces notes ont été estimées en exploitant les traces d'usage des utilisateurs, comme nous l'avions présenté précédemment (section 2.3.1, chapitre 2, partie 1). A partir de la matrice de similarité entre utilisateurs, les clusters sont construits sur la base des similarités de voisins, plutôt que des notes. Cette démarche utilisée pour le clustering a ainsi l'avantage de prendre également en compte des items non co-notés, vu que les similarités entre utilisateurs sont exploitées.

En vue d'illustrer cette démarche de clustering, nous proposons l'exemple très simple de la matrice de notes du tableau 2.1 qui représente cinq utilisateurs pouvant noter cinq items.
A partir des notes qu'ils ont attribué aux items, nous évaluons les similarités de notes entre ces utilisateurs (en utilisant le coefficient de Pearson). Le tableau 2.2 représente la matrice de similarité résultant de cette évaluation.
Par exemple, en considérant les items notés en commun, l'utilisateur u_1 et u_3 sont corrélés.

TAB. 2.1 – Matrice de note

	i_1	i_2	i_3	i_4	i_5
u_1	x		x		x
u_2	x	x			
u_3			x		
u_4	x				
u_5		x			

TAB. 2.2 – Matrice de similarité de note

	u_1	u_2	u_3	u_4	u_5
u_1		x	x		
u_2	x			x	x
u_3	x				
u_4		x			
u_5		x			

Bien évidemment, c'est la valeur de note (estimée) de u_1 et u_3 sur l'item co-noté i_3, qui détermine le degré de corrélation entre ces deux utilisateurs. Plus leur note sur l'item i_3 est similaire, plus ils sont corrélés, i.e. la valeur de similarité est proche de 1.

La matrice de similarité de notes va constituer, dans une étape suivante, les données d'entrée de l'algorithme de clustering PAM. La figure 2.2 décrit le processus du clustering PAM. Considérant que k, représentant le nombre de clusters à créer, est équivalent à 2, au début du processus deux médoïdes u_{med} et $u_{med'}$ (par exemple les utilisateurs u_3 et u_4) sont choisis aléatoirement (cf. figure 2.2 (1)). Ces médoïdes vont représenter les centres ou les médoïdes initiaux de chaque cluster. Par la suite, en calculant les dissimilarités (ou le coût de permutation) entre chacun de ces médoïdes et les autres utilisateurs (cf. figure 2.2 (2)), l'algorithme PAM identifie les médoïdes effectifs (par exemple les utilisateurs u_1 et u_2).
Il est à noter que cette opération itère jusqu'à ce que que les médoïdes deviennent stables, i.e., jusqu'à ce que les u_{med} et $u_{med'}$ ne changent plus (cf. section 1.3.3 du chapitre 1, partie 1).
A la fin du processus, nous obtenons deux clusters homogènes dont chacun comprend le groupe d'utilisateurs les plus similaires en terme de voisins (cf. figure 2.2 (3)). Selon l'exemple présenté, les deux clusters obtenus sont : $C_1 = \{u_1, u_3\}$ et $C_2 = \{u_2, u_4, u_5\}$. Nous pouvons constater par exemple que dans le cluster C_2, les utilisateurs partagent en effet des voisins communs, ce qui justifie leur appartenance au même cluster.

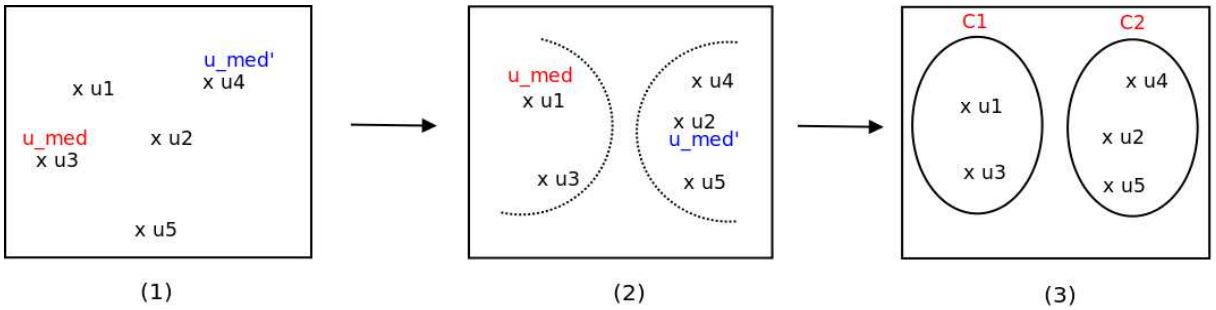


FIG. 2.2 – Clustering d'utilisateurs avec PAM

2.3 Calcul des similarités de comportement et génération des prédictions

Les similarités de comportement entre toute paire d'utilisateurs $\langle u_a, u_b \rangle$ sont évaluées au niveau de chaque cluster créé dans l'étape précédente. Cette évaluation repose sur l'algorithme d'extraction de motifs du BNCF et exploite l'équation (1.1) présentée dans le chapitre précédent (p. 77) pour l'évaluation des similarités.

De plus, dans l'objectif de réduire davantage l'espace de recherche des voisins et d'améliorer la qualité des prédictions, nous avons proposé de procéder à une sélection de sous-séquences positives qui comprennent uniquement les items positivement appréciés de la part des utilisateurs.
Si nous considérons une séquence β d'un utilisateur u_a, ω est une sous-séquence positive de β ($\omega \subset \beta$) lorsque tous les items contenus dans $\omega = \{i_1, i_2...., i_n\}$, ont des notes positives de u_a. Par exemple, sur une échelle de note $[1-5]$, nous pouvons considérer qu'une note v est positive lorsque $v >= 4$.

Ainsi, pour toute paire d'utilisateurs $\langle u_a, u_b \rangle$ appartenant à un même cluster, en prenant en considération les sous-séquences positives de u_a et u_b parmi leurs séquences de navigation, les motifs d'usage communs à ces deux utilisateurs sont extraits afin d'évaluer les similarités de comportement $SimNav(u_a, u_b)$.

Dès la génération de la matrice de similarité de comportement, comme dans le BNCF, la formule de prédiction basée sur la somme pondérée (cf. équation (1.2) p. 79) est employée en vue de calculer les prédictions. Cette étape correspond à la "PHASE II" du modèle BNCF-PCS (cf. figure 2.1).
Pour le calcul des prédictions, nous prenons en considération les plus proches voisins U_a (présents dans le même cluster que u_a) ayant déjà noté l'item et qui sont corrélés avec u_a.

2.4 Evaluation

Pour l'évaluation de la performance du BNCF-PCS, nous avons exploité le corpus d'usage du Crédit Agricole utilisé dans les expérimentations du chapitre précédent. Ce corpus comprend les traces d'usage correspondant aux activités de navigation des utilisateurs sur le portail Extranet.

Pour nos expérimentations, en considérant le nombre total d'utilisateurs présents dans le corpus et après avoir testé différents nombres de clusters à générer, nous avons choisi de générer 10 clusters afin d'obtenir des classes suffisamment denses, homogènes et représentatifs et éviter de dégrader la performance du système de recommandation à cause d'un éventuel manque de voisins au niveau des clusters.

Dans cette expérimentation, les différents modèles ont été évalués en terme de précision en utilisant les métriques MAE et HMAE ainsi qu'en terme de temps de calcul.

2.4.1 Modèles expérimentés

La précision des recommandations proposées par le BNCF-PCS a été comparée à différentes variantes des modèles BNCF et FCS. L'objectif de cette évaluation est d'examiner l'impact du clustering d'utilisateurs (algorithme PAM ou k-means) ainsi que l'influence de la nature de la matrice utilisée pour le clustering (matrice de note ou matrice de similarité). En outre, nous avons évalué l'impact de l'exploitation des sous-séquences positives sur le temps de calcul des similarités de comportement.

Notons que durant le calcul des prédictions pour le BNCF et le FCS, en considérant les résultats de l'expérimentation sur le corpus du Crédit Agricole (cf. section 1.3 du chapitre précédent, tableaux 1.4 et 1.6), nous avons fixé le minimum d'items co-notés à 20 et le seuil θ à 0.2.

2.4.2 Résultats

MAE

Dans cette section, nous présentons les résultats d'évaluation en terme de MAE. Dans cette évaluation, il était question d'examiner d'abord l'impact du clustering exploitant une matrice de notes dans le cadre du BNCF et du FCS. Par la suite, l'objectif était d'évaluer le BNCF-PCS en examinant l'impact de l'algorithme de clustering exploitant une matrice de similarité.

Le tableau 2.3 présente les résultats en terme de MAE avec l'application ou non d'un clustering d'utilisateurs exploitant la matrice de note, dans le cadre du BNCF ou du FCS. Il est à signaler que les sous-séquences positives ne sont pas utilisées dans ce cadre.
En observant les résultats de ce tableau, nous pouvons d'abord remarquer que sans l'application du clustering, la précision est légèrement plus faible dans le cas du BNCF, comparé au FCS (cf. section 1.3.1 du chapitre précédent).
Nous remarquons en outre que l'application du clustering d'utilisateurs au FCS à partir de la matrice de note (correspondant au FCS basé sur le clustering dans la figure 2.1), engendre une dégradation de la précision des recommandantions. Ceci peut découler du fait que le clustering appliqué dans ce cas repose uniquement sur les items co-notés pour grouper les utilisateurs en clusters, ce qui risque de négliger certains voisins pertinents.
Ces résultats confirment l'état de l'art. En effet, l'étude réalisée par [Sarwar *et al.*, 2002] a montré que, malgré son intérêt pour le passage à l'échelle dans le cadre du FCS, le clustering d'utilisateurs (basé sur k-means et exploitant une matrice de notes) a un impact sur la performance du système de recommandation dans la mesure où la précision des prédictions tend à être faible.

En outre, à partir du tableau 2.3, nous constatons que le clustering PAM mène à la plus faible précision (baisse de 5% de précision comparée à la précision du FCS sans clustering). Dans le cas du BNCF, l'utilisation du clustering PAM à partir d'une matrice de notes entraîne également une dégradation de la précision, ce qui est similaire aux résultats du FCS.

Tab. 2.3 – Résultats en MAE avec et sans clustering (utilisation d'une matrice de note en cas de clustering)

	FCS	**BNCF**
Sans clustering	**0.763**	0.789
Avec clustering k-means	0.782	0.797
Avec clustering PAM	0.799	0.825

Le modèle BNCF-PCS applique un clustering exploitant une matrice de similarité. Par la suite, les sous-séquences positives des utilisateurs appartenant aux mêmes clusters créés, sont considérées en vue d'évaluer les similarités de comportement.
Dans le but d'examiner la performance du clustering utilisé par le BNCF-PCS, nous présentons dans le tableau 2.4, les résultats en MAE en cas d'application de l'algorithme PAM, comparé à l'algorithme k-means.

Tab. 2.4 – Résultats en MAE : utilisation d'une matrice de similarité pour le clustering

BNCF-PCS	
Avec clustering k-means	0.780
Avec clustering PAM	**0.674**

Selon les résultats du tableau 2.4, nous constatons que l'application du clustering exploitant une matrice de similarité d'utilisateurs contribue à une amélioration de la

MAE, quel que soit l'algorithme de clustering utilisé, comparée aux résultats du tableau 2.3 relatifs au BNCF. De plus, dans le cas d'un clustering PAM (BNCF-PCS), la précision atteint même une amélioration de 15%, par rapport à un clustering exploitant une matrice de note.
Rappelons qu'ici le clustering a été appliqué à une matrice de similarité, ce qui permet de générer des clusters, non pas uniquement en fonction de la manière dont les utilisateurs ont co-noté les items, mais également suivant les similarités de voisins que ces utilisateurs ont en commun. En outre, cette démarche de clustering ne considère pas seulement les items co-notés, mais l'ensemble des items notés par les utilisateurs. Il semblerait que dans cette expérimentation, la considération des voisinages communs lors du clustering PAM, contribue à l'amélioration de la performance du système.
De plus, lors de l'évaluation des similarités, le BNCF-PCS exploite l'information relative aux items positivement appréciés par les utilisateurs, contenus dans les sous-séquences positives. Les résultats du tableau 2.4 confirment également que cette stratégie permet d'améliorer le calcul des voisinages et l'identification des plus proches voisins, ce qui mène à une meilleure qualité des recommandations en terme de MAE.

Nous pouvons ainsi déduire que l'amélioration de la précision des prédictions (en MAE) résulte de l'application de l'algorithme PAM sur une matrice de similarité et de l'utilisation des sous-séquences positives des utilisateurs pour l'évaluation des similarités de comportement.

HMAE

Comme dans le chapitre précédent, nous nous intéressons ici à l'évaluation de la HMAE du BNCF-PCS tout en comparant les mêmes variantes utilisées ci-dessus, i.e. avec ou sans clustering, utilisation d'une matrice de note ou bien d'une matrice de similarité. Les résultats en HMAE sont présentés dans les tableaux 2.5 et 2.6.

A partir du tableau 2.5, nous observons que lorsque le clustering exploite la matrice de note, les valeurs de la HMAE augmentent pour les deux modèles BNCF et FCS. Or, sans l'utilisation du clustering, le BNCF atteint une meilleure performance (amélioration d'environ 7%) en HMAE, comparé au FCS (cf. section 1.3.1 du chapitre précédent).

Tab. 2.5 – Résultats en HMAE avec ou sans clustering (utilisation d'une matrice de note en cas de clustering)

	FCS	**BNCF**
Sans clustering	0.541	**0.501**
Avec clustering k-means	1.285	1.272
Avec clustering PAM	1.168	1.159

Le tableau 2.6 présente les résultats en HMAE, en cas d'application du clustering exploitant une matrice de similarité, dans le cadre du BNCF-PCS.

Lorsque cette matrice est utilisée, la HMAE baisse considérablement pour les deux algorithmes de clustering k-means et PAM, comparée à la HMAE obtenue lorsque le clustering exploite une matrice de notes (cf. tableau 2.5).
Bien qu'il n'y a pas d'amélioration de résultats de la HMAE, comparé à la variante "sans clustering", rappelons qu'une amélioration importante est obtenue au niveau de la MAE et, comme nous le préciserons dans ce qui suit, le temps de calcul des voisinages a été réduit.

TAB. 2.6 – Résultats en HMAE : utilisation d'une matrice de similarité pour le clustering

BNCF-PCS	
Avec clustering k-means	**0.587**
Avec clustering PAM	0.603

Temps de calcul

Dans cette section, nous nous intéressons à l'évaluation du temps de calcul requis pour la phase de calcul des similarités de comportement, avec ou sans clustering et avec la sélection ou non des sous-séquences positives.

Les résultats de cette évaluation ont montré que les modèles n'intégrant pas le clustering, requièrent en moyenne un temps de calcul plus élevé, en vue d'évaluer les similarités de comportement. Ce temps de calcul résulte du fait que les similarités ont été évaluées entre toutes les paires d'utilisateurs contenues dans le corpus d'apprentissage. Or, avec l'application du clustering, ces similarités sont calculées uniquement au sein des clusters, ce qui se répercute sur le nombre d'utilisateurs concernés par l'évaluation, qui tend bien évidemment à la baisse.

Par ailleurs, à partir de ces résultats, nous remarquons également que la sélection des sous-séquences positives contribue à une importante réduction du temps de calcul. En effet, ce temps décroît d'environ 8% sans l'utilisation du clustering et de 16% à 30% avec l'application du clustering. Cela peut être expliqué par le fait que le nombre de séquences considérées lors de l'extraction des motifs d'usage a été réduit.
Pour le BNCF-PCS, l'application du clustering et l'utilisation des sous-séquences positives reste bénéfique en terme de temps de calcul ainsi qu'en terme de précision des recommandations.

2.4.3 Discussion

Nous avons proposé le modèle BNCF-PCS en vue de réduire l'espace de recherche pour l'identification de voisins et d'améliorer la performance du système de recommandation.

Pour la réduction de l'espace de recherche, le BNCF-PCS applique l'algorithme de clustering PAM. La particularité de ce clustering réside dans l'utilisation d'une matrice de similarité "Utilisateur x Utilisateur" plutôt qu'une matrice de note "Utilisateur x Item", afin de créer des clusters. Ainsi, dans le cadre du BNCF-PCS, les utilisateurs sont groupés en différents clusters homogènes, selon les similarités de leurs voisins.
L'avantage d'une telle démarche de clustering est la considération d'items supplémentaires et non pas uniquement des items co-notés par les utilisateurs. En effet, étant donné que les similarités exploitées pour le clustering reposent sur les voisinages communs, tous les items consultés par les utilisateurs ayant des voisins en commun sont considérés.

Le BNCF-PCS a été évalué en termes de MAE et de HMAE et comparé à d'autres modèles de FCS, en vue d'examiner l'influence de la matrice utilisée lors du clustering ainsi que l'impact de l'algorithme de clustering utilisé.
Les résultats montrent l'intérêt d'appliquer le clustering PAM (exploitant une matrice de similarité) et d'utiliser les sous-séquences positives pour évaluer les similarités de comportement. En effet, une importante amélioration en terme de MAE a été atteinte (cf. tableau 2.4).

Toutefois, avec la sélection des sous-séquences positives, le système risque de ne pas tenir compte d'informations pertinentes relatives aux séquences utilisateurs, en vue de détecter des motifs d'usage fiables. En effet, des sous-séquences incluant les items non appréciés peuvent également révéler certaines corrélations de comportement entre utilisateurs.

Par ailleurs, l'application du clustering (sur une matrice de similarité) risque de négliger certaines informations pertinentes pendant le processus de réduction de l'espace de recherche. En effet, les utilisateurs sont groupés en clusters selon les similarités de voisins. De ce fait, deux utilisateurs u_a et u_b qui sont faiblement similaires avec leur voisin commun u_c, ne vont pas appartenir au même cluster. Néanmoins, l'utilisateur u_b peut apporter une importante contribution à la génération de prédictions à l'utilisateur actif u_a, surtout lorsque le système ne retrouve pas d'autres voisins à u_a. Par conséquent, une telle perte d'information est susceptible d'engendrer une diminution de la capacité prédictive du système de recommandation.

C'est dans ce contexte que nous avons proposé d'étendre notre approche de recommandation, en intégrant d'autres techniques permettant de faire face à ce problème de perte d'information. En effet, nous souhaitons améliorer le processus d'identification des voisins, notamment par la découverte de nouveaux liens entre utilisateurs qui peuvent être interprétés comme étant des similarités. Ces nouveaux liens représentent une solution prometteuse face au problème de manque de données.

Troisième partie

Approche sociale de recommandation

Chapitre 1

Prédiction de lien dans les réseaux comportementaux

Dans l'objectif de pallier le manque de données, d'identifier des voisins fiables et de promouvoir la performance des systèmes de recommandation, nous avons proposé une nouvelle approche sociale de recommandation.
En effet, dans le cadre du FCS (Filtrage Collaboratif Standard), les voisins sont identifiés sur la base des similarités entre un utilisateur actif et les autres utilisateurs. L'évaluation de ces similarités repose sur le calcul des corrélations de leurs appréciations vis-à-vis d'items co-notés dans le passé. Or, l'inconvénient de cette approche est qu'elle exploite uniquement les appréciations communes, i.e. les liens directs entre utilisateurs, afin de calculer les prédictions. En effet, si deux utilisateurs ne partagent aucune de ces appréciations communes, aucun lien ne peut être établi entre eux (ce lien est même considéré comme nul).
Ce problème émane notamment du manque de données. En effet, lorsque le volume des données de notes est limité, l'identification des voisins s'avère complexe, ce qui entraîne une diminution de la capacité prédictive et de la qualité des prédictions produites par le système de recommandation.

Dans la partie précédente, nous avons déjà fait une première proposition pour pallier le problème de manque de données. L'approche de recommandation présentée dans ce chapitre vise également à remédier à ce problème. En effet, l'objectif de cette nouvelle approche est d'explorer de nouveaux liens entre des utilisateurs n'ayant pas eu nécessairement des appréciations communes antérieurement. C'est dans cette optique que nous nous sommes inspirés des approches issues de l'analyse des réseaux sociaux, permettant notamment de prédire les liens entre utilisateurs, d'où l'appellation d'approche sociale de recommandation.

Les réseaux sociaux représentent une structure sociale entre des acteurs, souvent des individus ou des organisations, permettant d'indiquer les connexions existantes entre eux,

au travers de divers liens sociaux tels que l'amitié, la collaboration professionnelle ou bien l'échange d'information [Jamali et Abolhassani, 2006].
Avec l'évolution accrue du Web et notamment du Web social, l'analyse des réseaux sociaux est de plus en plus prépondérante. Elle permet en effet d'analyser les interactions, d'examiner leur évolution et de comprendre les flux sociaux. En vue d'analyser l'évolution de ces interactions, diverses techniques peuvent être utilisées, dont notamment la prédiction de lien [Liben-Nowell et Kleinberg, 2003]. L'objectif de la prédiction de lien consiste à prédire les futures interactions entre les acteurs, i.e. les futurs liens qui vont potentiellement apparaître dans un réseau social.

Du point de vue modélisation de relations entre acteurs ou utilisateurs, les méthodes de prédiction de lien et les systèmes de recommandation convergent vers une même question de recherche : comment identifier de nouveaux liens ou relations entre des utilisateurs qui ne sont pas reliés (connectés) ?

Dans ce contexte, nous avons proposé le modèle D-BNCF "Densified-Behavioral Network based Collaborative Filtering" [Esslimani *et al.*, 2009b] [Esslimani *et al.*, 2009c] [Esslimani *et al.*, 2010a]. En se basant sur les similarités comportementales calculées par le BNCF, le modèle D-BNCF modélise les liens entre utilisateurs au travers d'un réseau comportemental. Ainsi, deux utilisateurs similaires sont reliés dans ce réseau.
D-BNCF exploite par la suite les méthodes de prédiction de lien, notamment les associations transitives, en vue de découvrir de nouveaux liens reliant les utilisateurs. Notre choix d'intégrer les méthodes de prédiction de lien dans le cadre du D-BNCF, a été appuyé par leur succès dans le contexte des réseaux sociaux [Liben-Nowell et Kleinberg, 2003].
Dans le domaine des réseaux sociaux, les associations transitives signifient que "les amis de mes amis, sont mes amis". La transposition de cette propriété dans un réseau comportemental implique que "les utilisateurs qui se comportent comme ceux qui se comportent comme moi, se comportent comme moi".
L'application des méthodes de prédiction de lien, intégrant la transitivité, permet de compléter et de densifier le réseau comportemental par de nouveaux liens. Ces nouveaux liens sont interprétés comme des relations entre utilisateurs. Ainsi, de nouveaux voisins sont identifiés et intégrés au calcul des recommandations.

1.1 Prédiction de lien

1.1.1 Dans le domaine des réseaux sociaux

Au vu du succès et de la popularité croissante des réseaux sociaux, l'analyse de ces derniers a sucité l'intérêt d'innombrables travaux de recherche [Barabási *et al.*, 2002], [Liben-Nowell et Kleinberg, 2003], [Mislove *et al.*, 2007], [Crandall *et al.*, 2008]. La plupart de ces travaux visent à analyser les structures des réseaux sociaux afin de représenter les

interactions et les influences entre les acteurs.

La prédiction de lien constitue l'un des problèmes majeurs dans le domaine de l'analyse des réseaux sociaux. Elle consiste à analyser l'évolution d'un réseau en prédisant les futurs liens qui seront rajoutés et en inférant les interactions futures entre les nœuds de ce réseau.
La prédiction de lien a été étudiée dans différents types de réseaux sociaux tels que les réseaux scientifiques de co-auteurs [Newman, 2001], [Barabási *et al.*, 2002], les réseaux d'interactions biologiques [Yamanishi *et al.*, 2005], [Ohn *et al.*, 2003], les réseaux de communication via les forums ou par e-mail [Lim *et al.*, 2003], etc.

Les méthodes de prédiction de lien peuvent reposer sur des approches d'analyse topologique ou bien d'analyse d'attributs [Cooke, 2006].
Les approches d'analyse topologique utilisent uniquement le réseau (représenté par un graphe) afin d'inférer les futures interactions (deux personnes ayant des amis en commun, ont tendance à interagir). Une étude comparative de cette classe de méthodes est présentée dans [Liben-Nowell et Kleinberg, 2003].
Les approches d'analyse d'attributs n'intègrent pas de théorie de graphe, mais considèrent plutôt le contenu des interactions entre les nœuds (les personnes) ou leurs attributs. A titre d'exemple, ces approches peuvent considérer le contenu des communications entre les personnes en vue de rechercher les intérêts qu'ils ont en commun (deux personnes qui discutent de la plongée sous-marine et de l'algèbre ont tendance à partager les mêmes centres d'intérêts et peuvent ainsi être reliés).
Par ailleurs, [Bartal *et al.*, 2009] montrent l'intérêt de prédire les liens en combinant ces deux classes de méthodes dans le cadre d'un réseau de collaboration scientifique. Il s'agit des méthodes de prédiction de lien exploitant les approches topologiques et des méthodes issues des approches d'analyse d'attributs basées sur l'analyse de contenu. Les critères de collaboration scientifique ("co-authoring" de publications scientifiques) et de similarité des thématiques de recherche sont considérés pour la prédiction de lien.

1.1.2 Dans le domaine des systèmes de recommandation

Dans le domaine des systèmes de recommandation, les méthodes de prédiction de lien peuvent être utilisées pour prédire de nouveaux liens (entre utilisateurs et/ou items) et générer les recommandations. [Huang *et al.*, 2002] proposent un système de recommandation basé sur un graphe bi-partite dans le contexte des bibliothèques électroniques. Les items (les livres) et les utilisateurs représentent les nœuds, les arcs reliant les utilisateurs aux items représentent les transactions. Des associations dites de "haut-degré" (exploitant l'algorithme Hopfield issu du domaine des réseaux de neurones [Hopfield, 1982]) ont été appliquées afin de rechercher des liens pouvant relier des nœuds non connectés (représentant les utilisateurs et les items).

[Papagelis *et al.*, 2005] exploitent une méthode de prédiction de lien fondée sur la

transitivité, en vue de remédier au problème de manque de données. Un modèle exploitant les inférences de confiance est proposé, dans le but d'augmenter les voisinages requis pour la génération des recommandations. Dans la même optique, [Huang *et al.*, 2005] exploitent les méthodes de prédiction de lien basées sur les voisins et sur les chemins dans le but d'analyser les interactions utilisateur-item représentées à travers un graphe bi-partite.

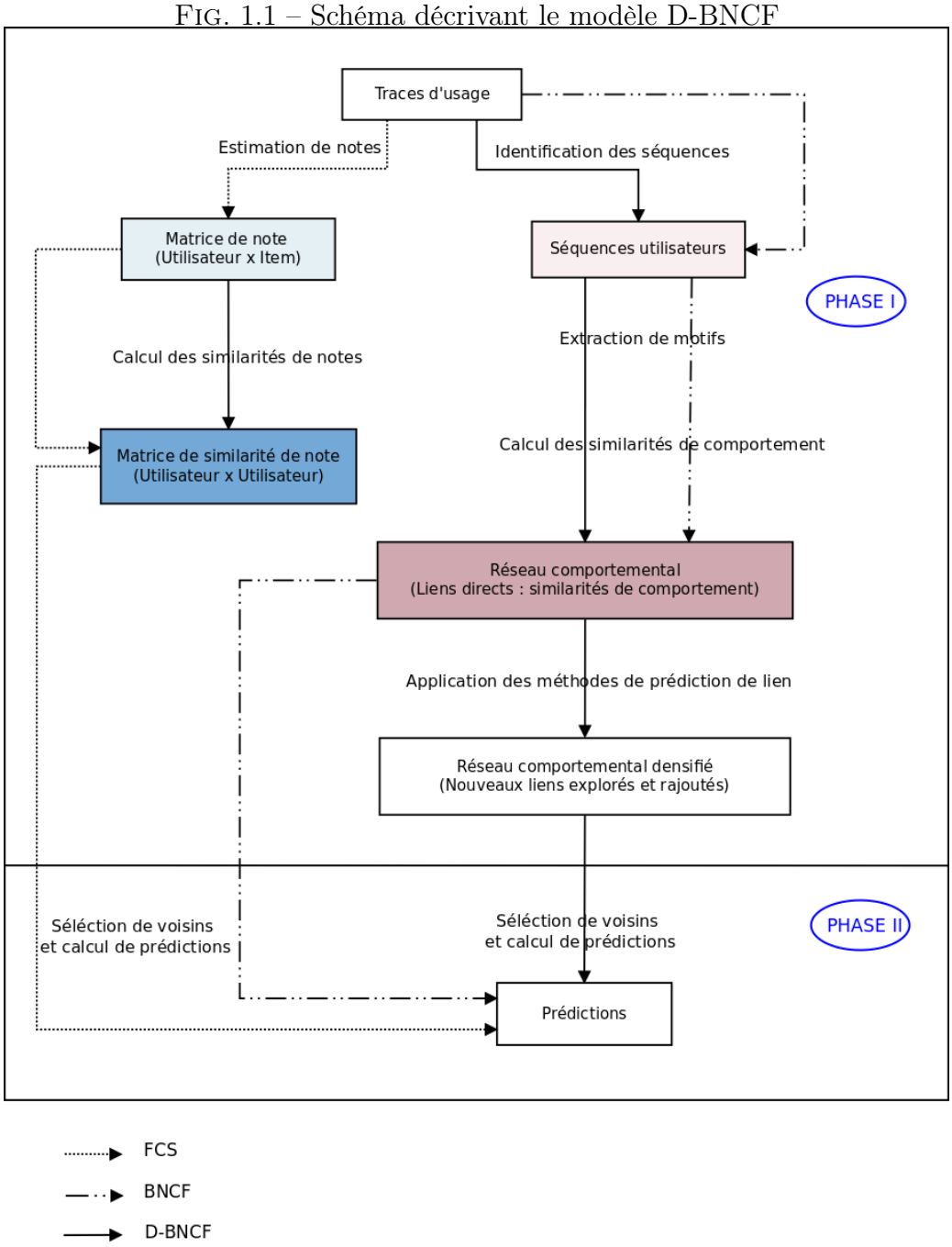
Par ailleurs, [Kautz *et al.*, 1997] [Zheng *et al.*, 2007] intègrent l'information sociale dans le cadre du FCS. Les consommateurs ou les utilisateurs représentent les nœuds et les relations sociales représentent les liens. Dans le but d'identifier les voisins, [Zheng *et al.*, 2007] appliquent une méthode de prédiction de lien fondée sur le calcul des distances (les plus courts chemins) entre utilisateurs au niveau du réseau social. Les nouveaux liens sociaux découverts sont utilisés pour calculer les prédictions.

A la différence des travaux présentés ici, dans le cadre du D-BNCF nous proposons d'exploiter les méthodes de prédiction de lien dans un réseau comportemental. Ce réseau comprend un seul type de nœud représentant les utilisateurs. Les liens reliant ces utilisateurs reposent sur l'information comportementale plutôt que l'information sociale ou transactionnelle considérée souvent par la plupart des études, notamment celles présentées ici.
L'objectif de l'application des méthodes de prédiction de lien est d'identifier de nouveaux liens entre utilisateurs. Ces nouveaux liens seront intégrés dans le processus de recommandation afin de pallier le problème de manque de données et d'améliorer la précision des prédictions.
Le point commun entre le D-BNCF et les travaux cités ci-dessus, réside dans la considération du principe d'associations transitives pour la découverte de nouveaux liens.

1.2 Modèle D-BNCF

La figure 1.1 présente le modèle D-BNCF en comparaison au BNCF et au FC Standard (FCS). Pour rappel, le FCS et le BNCF visent à identifier les liens ou les voisins directs en exploitant respectivement les similarités de notes et de comportement entre utilisateurs (PHASE I). Ces similarités sont utilisées par la suite pour l'identification des plus proches voisins dont les appréciations sont combinées pour le calcul des prédictions (PHASE II).
Quant au D-BNCF, il exploite la démarche utilisée par le BNCF consistant à l'évaluation des similarités comportementales à partir des motifs d'usage communs. Ces similarités permettent de modéliser les liens entre utilisateurs à travers un réseau comportemental. Le D-BNCF inclut une étape supplémentaire permettant de densifier ce réseau comportemental grâce à l'intégration de nouveaux liens (voisins) identifiés par les méthodes de prédiction de lien. Ces liens sont impliqués par la suite dans le calcul des prédictions (PHASE II).
Les différents méchanismes utilisés par le D-BNCF vont être explicités dans les sous-sections suivantes.

FIG. 1.1 – Schéma décrivant le modèle D-BNCF



1.2.1 Modélisation du réseau comportemental

Dans le cadre du D-BNCF, nous exploitons un réseau d'utilisateurs. A la différence des réseaux sociaux qui reposent sur les liens sociaux, ce réseau utilise l'information comportementale pour établir des liens entre des utilisateurs ayant des motifs d'usage en commun. Le D-BNCF exploite la démarche du BNCF (en phase d'apprentissage) permettant d'évaluer les similarités de comportement entre les paires d'utilisateurs sur la base des motifs

d'usage communs, en vue de construire un réseau comportemental. Ce réseau est modélisé à travers un graphe où les nœuds représentent les utilisateurs, les arcs représentent les liens entre eux et les similarités de comportement représentent les poids des arcs.

1.2.2 Densification du réseau comportemental

En vue de densifier le réseau comportemental construit, nous exploitons les méthodes de prédiction de lien topologiques basées sur les voisinages des nœuds et sur les chemins [Liben-Nowell et Kleinberg, 2003] [Adamic et Adar, 2003] [Newman, 2001].
Nous n'avons pas choisi d'appliquer au D-BNCF l'approche d'analyse d'attribut (ou de contenu) parce qu'elle risque de ne pas prédire certains liens entre utilisateurs, en raison de la non similarité des contenus des items préalablement visités (e.g. si deux utilisateurs consultent des items ayant un contenu différent, cette approche n'établit pas de lien entre eux). Par conséquent, le nombre de liens prédits par l'approche d'analyse d'attribut risque d'être faible.

En outre, dans le cadre du D-BNCF exploitant les méthodes de prédiction de lien topologiques, les liens entre deux utilisateurs sont calculés, non pas en considérant uniquement les items consultés ou notés en commun, mais potentiellement tous les items que ces utilisateurs ont déjà consulté. Notons que les liens calculés ne dépendent pas que de ces deux utilisateurs, mais d'autres informations obtenues à partir du réseau ou du graphe.
Ce principe de calcul de liens rejoint la démarche de clustering (exploitant une matrice de similarité) proposée dans le chapitre précédent, dans la mesure où elle consiste à prendre en compte les autres utilisateurs n'ayant pas nécessairement des items en commun avec l'utilisateur actif.
Les méthodes de prédiction de lien que nous avons utilisées dans le cadre du D-BNCF sont présentées dans ce qui suit.

Méthodes basées sur le voisinage

- **Attachement préférentiel** : [Barabási *et al.*, 2002] et [Newman, 2001] considèrent qu'il existe une forte probabilité que des nœuds se connectent, si ces nœuds, appellés également "hubs", sont déjà connectés à un nombre élevé de nœuds à travers le réseau. Cette idée rejoint le principe du "rich-get-richer"[22].
  Selon [Liben-Nowell et Kleinberg, 2003], *l'attachement préférentiel* peut être mesuré comme étant la probabilité de connexion entre deux nœuds u_a et u_b basée sur le produit du nombre de leurs voisins.

[22] A l'origine, ce principe a été utilisé dans le domaine de l'économie pour critiquer le capitalisme et en particulier le fait que les personnes riches ont tendance à s'enrichir plus. Il a été repris par la suite par Barabási en 1999 qui a constaté que ce principe est valable également pour prédire l'évolution des liens hypertextes entre les pages Web [Barabási et Albert, 1999]

L'attachement préférentiel a toutefois l'inconvénient d'obtenir des valeurs de similarités élevées concernant les utilisateurs hyperconnectés, au détriment des utilisateurs peu connectés dans le réseau. Cet inconvénient relève du fait que les relations entre utilisateurs dépendent uniquement de leur connectivité. Or, notre but est de trouver de nouveaux voisins aux nœuds qui en ont peu.
En outre, une autre limite de cette méthode est la création de plusieurs liens entre les nœuds (ayant des nœuds voisins) et la maximisation de la connectivité du réseau. Donc cette méthode s'avère peu appropriée dans notre cas.
Nous considérons $\Gamma_{(u_a)}$ qui représente les voisins de l'utilisateur u_a et $\left|\Gamma_{(u_a)}\right|$ le nombre de voisins de u_a. *L'attachement préférentiel* entre u_a et u_b est calculé comme suit :

$$Sim(u_a, u_b) = \frac{1}{\xi}(\left|\Gamma_{(u_a)}\right| \cdot \left|\Gamma_{(u_b)}\right|) \tag{1.1}$$

ξ représente ici un facteur de normalisation.

– **Voisins communs "Common neighbors"** : mesure la similarité entre deux utilisateurs u_a et u_b en fonction du nombre de leurs voisins communs. *Voisins communs* entre u_a et u_b est calculé ainsi :

$$Sim(u_a, u_b) = \frac{1}{\xi}(\left|\Gamma_{(u_a)} \cap \Gamma_{(u_b)}\right|) \tag{1.2}$$

ξ représente également ici un facteur de normalisation.
Cette méthode considère que plus les utilisateurs partagent des voisins en commun, plus ils sont corrélés [Liben-Nowell et Kleinberg, 2003]. Or, comme pour *l'attachement préférentiel*, l'inconvénient de cette méthode est sa tendance à attribuer des similarités élevées aux utilisateurs ayant de nombreux voisins. De ce fait, la similarité entre les utilisateurs disposant de peu de voisins tend à être faible, voire nulle, alors que notre objectif initial consiste à créer des liens, en particulier pour cette catégorie d'utilisateurs.

– **Coefficient Jaccard** : il s'agit d'une amélioration de la méthode *voisins communs*, puisqu'elle évalue la similarité comme étant le rapport entre les voisins communs de u_a et u_b et le nombre total de leurs voisins [Liben-Nowell et Kleinberg, 2003]. Selon l'équation (1.3), plus u_a et u_b ont des voisins communs, parmi l'ensemble de leurs voisins, plus ils sont corrélés.
Comparé aux deux méthodes précédentes, *Jaccard* a l'avantage de ne pas augmenter l'influence des utilisateurs disposant d'un grand nombre de voisins.

$$Sim(u_a, u_b) = \frac{\left|\Gamma_{(u_a)} \cap \Gamma_{(u_b)}\right|}{\left|\Gamma_{(u_a)} \cup \Gamma_{(u_b)}\right|} \tag{1.3}$$

– **Adamic/Adar** : à l'origine, [Adamic et Adar, 2003] ont proposé une méthode afin d'évaluer la probabilité qu'un utilisateur u_a soit connecté à u_b, en prenant en compte

les items que ces deux utilisateurs ont en commun. La particularité de cette méthode est que les items qui sont partagés par peu d'utilisateurs, ont un poids plus important que les items dont les occurences sont élevées (i.e. les items qui sont communs à plusieurs paires d'utilisateurs).
Selon l'équation (1.4), au lieu de considérer les items, nous considérons les voisins que u_a et u_b ont en commun. La fréquence de chaque voisin commun u_c, noté $frequency(u_c)$, est calculée parmi toutes les paires d'utilisateurs.
L'avantage de cette méthode est qu'elle met en évidence l'importance des voisins communs qui sont rares.

$$Sim(u_a, u_b) = \sum_{u_c \in \Gamma_{(u_a)} \cap \Gamma_{(u_b)}} \frac{1}{\log\left[frequency(u_c)\right]} \tag{1.4}$$

– Outre les méthodes présentées ci-dessus, nous avons proposé une nouvelle méthode de prédiction de lien, fondée sur le voisinage, appelée "**ETL (Enhanced Transitive Link)**". Il s'agit d'une amélioration de la méthode *Jaccard*. Cette méthode, représentée par l'équation (1.5), calcule le lien entre deux utilisateurs en considérant les plus proches voisins que deux utilisateurs u_a et u_b ont en commun, notés $\left|E_{(u_a)} \cap E_{(u_b)}\right|$, par rapport à leurs voisins communs, notés $\left|\Gamma_{(u_a)} \cap \Gamma_{(u_b)}\right|$.
Pour séléctionner les plus proches voisins communs à u_a et u_b, nous avons proposé de calculer pour chaque utilisateur, la valeur médiane de similarité, parmi l'ensemble de ses voisins. Ainsi, les plus proches voisins de chaque utilisateur sont déterminés en fonction de cette valeur médiane de similarité.

$$Sim(u_a, u_b) = \frac{\left|E_{(u_a)} \cap E_{(u_b)}\right|}{\left|\Gamma_{(u_a)} \cap \Gamma_{(u_b)}\right|} \tag{1.5}$$

Lors de l'application des méthodes fondées sur le voisinage, une seule itération est permise afin d'explorer de nouveaux liens à travers le réseau comportemental. Les liens originaux sont ainsi remplacés par les nouveaux liens calculés, qui sont intégrés par la suite pour générer les prédictions.

Méthodes basées sur les chemins

Distance de graphe ("graph distance") : dans le but de comparer les méthodes précédentes appartenant à la famille de méthode de prédiction basées sur le voisinage, à la famille de méthode basées sur les chemins [Cooke, 2006], nous avons utilisé la méthode *distance de graphe* pour identifier les nouveaux liens à travers le réseau comportemental. Cette méthode calcule le plus court chemin entre les utilisateurs u_a et u_b. Dans notre modèle, nous avons calculé les plus courts chemins en prenant en compte les similarités de comportement comme étant les poids des arcs.
Nous avons transformé les similarités entre les utilisateurs u_a et u_b, notées $Sim(u_a, u_b)$, en valeurs de distance $d(u_a, u_b)$ selon l'équation (1.6).

$$d(u_a, u_b) = 1 - Sim(u_a, u_b) \tag{1.6}$$

A la fin du processus, lorsque les nouveaux liens sont découverts (les plus courts chemins entre utilisateurs), ces valeurs sont à nouveau transformées en similarités $Sim'(u_a, u_b)$ (suivant l'intervalle $[0, 1]$) en utilisant l'équation (1.7). d_{max} représente la distance maximale d'un plus court chemin identifié parmi toutes les paires d'utilisateurs. L'objectif de l'utilisation de cette équation consiste à normaliser les valeurs de distances.
Ainsi, les nouveaux liens calculés sont considérés pour la génération des prédictions.

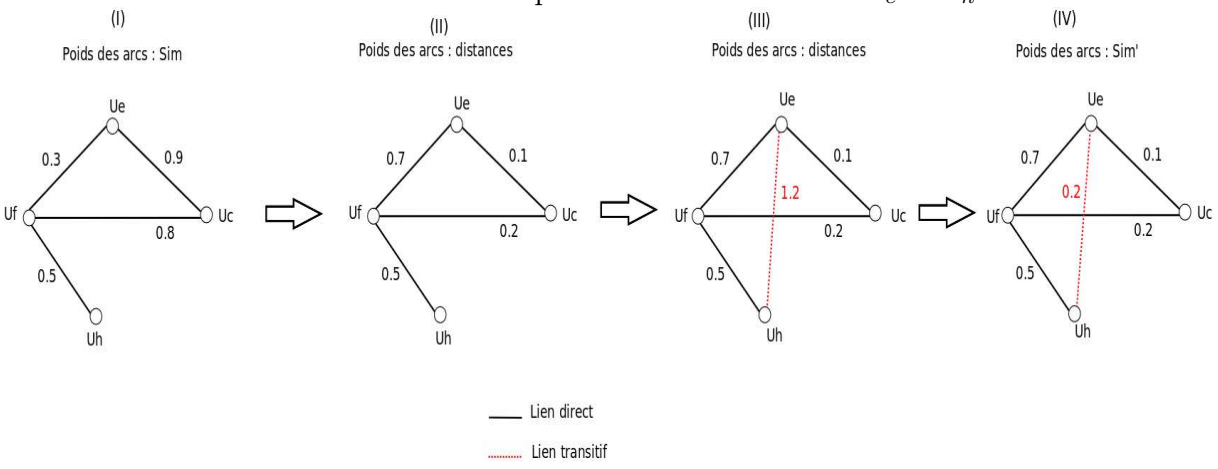
$$Sim'(u_a, u_b) = 1 - \frac{d(u_a, u_b)}{d_{max}} \tag{1.7}$$

Les figures 1.2 et 1.3 permettent d'illustrer l'application de cette méthode. Lors de la recherche des plus courts chemins entre les paires d'utilisateurs à travers le réseau comportemental, nous distinguons deux types de paires :

1. une paire d'utilisateurs qui ne sont pas connectés directement,
2. une paire d'utilisateurs qui sont déjà connectés à travers un lien direct.

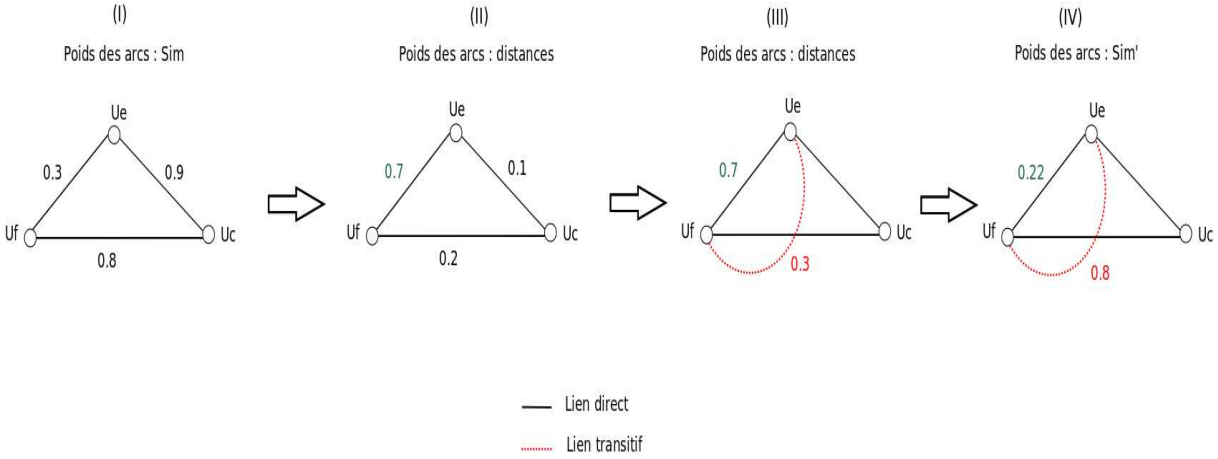
La figure 1.2 est une illustration du premier type de paire. Dans cet exemple, u_e et u_h ne sont pas connectés directement dans le réseau comportemental. Les similarités sont transformées en distance pour que les chemins les plus courts soient calculés à travers le réseau comportemental (volet (II)). Grâce à la transitivité, un nouveau lien est identifié entre u_e et u_h tel que : $d(u_e, u_h) = d(u_e, u_f) + d(u_f, u_h) = 0.5 + 0.7 = 1.2$ (volet (III)). Dans cet exemple, nous avons considéré que $d_{max} = 1.5$. Ainsi la similarité est calculée en utilisant l'équation (1.7) : $Sim'(u_e, u_h) = 1 - \frac{d(u_e, u_h)}{d_{max}} = 1 - \frac{1.2}{1.5} = 0.2$ (volet (IV)).

FIG. 1.2 – Calcul du plus court chemin entre u_e et u_h



L'application de la méthode *distance de graphe* sur le deuxième type de paire, permet d'explorer de nouveaux liens potentiellement forts entre deux utilisateurs reliés, au vu des liens forts reliant les voisins intermédiaires.
La figure 1.3 est une illustration de ce cas. Ainsi, en considérant les deux utilisateurs u_e et u_f tel que $Sim(u_e, u_f) = 0.3$, lors du calcul des plus courts chemins à travers le réseau comportemental, toutes les similarités sont transformées en distance : $d(u_e, u_f) = 1 - Sim(u_e, u_f) = 1 - 0.3 = 0.7$ (volet (II)). Par la suite, de nouveaux liens sont identifiés grâce à la transitivité. Ici, nous découvrons un deuxième chemin ou lien entre u_e et u_f à travers l'utilisateur u_c. Une nouvelle distance peut être ainsi calculée en prenant en compte ce nouveau lien (volet (III)) :
$d'(u_e, u_f) = d(u_e, u_c) + d(u_c, u_f) = 0.1 + 0.2 = 0.3$
Alors la nouvelle similarité est calculée en utilisant $d'(u_e, u_f)$ comme valeur de distance dans l'équation (1.7). Dans cet exemple, nous avons considéré également que $d_{max} = 1.5$, de ce fait, la similarité $Sim'(u_e, u_f) = 1 - \frac{d'(u_e,u_f)}{d_{max}} = 1 - \frac{0.3}{1.5} = 0.8$ (volet (IV)). Ainsi, en prenant en compte cette nouvelle similarité calculée, le nouveau lien remplace l'ancien lien qui reliait les deux utilisateurs.

FIG. 1.3 – Calcul du plus court chemin entre u_e et u_f

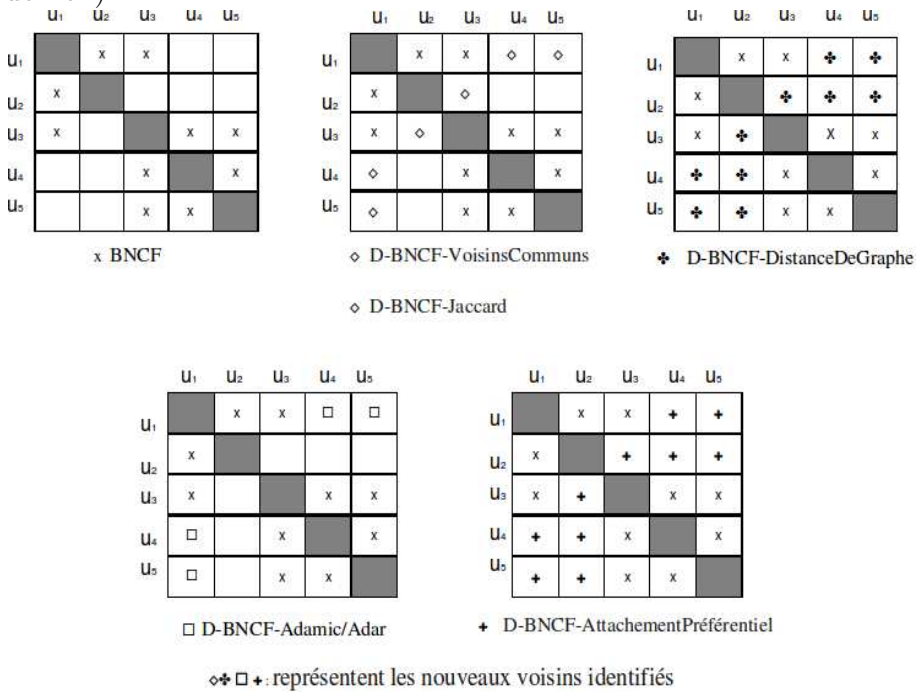


Dans le cadre des méthodes fondées sur les chemins, nous avons suggéré une variation de la méthode *distance de graphe*, en considérant le critère du nombre de nœuds intermédiaires présents au niveau du plus court chemin entre deux paires de nœuds ou d'utilisateurs dans le réseau comportemental. Ainsi, les poids des arcs, définis par les similarités de comportement, ne sont pas considérés. Plus le nombre de nœuds intermédiaires est faible, plus le chemin est fiable.
Dans le but de calculer le nouveau lien entre deux utilisateurs u_a et u_b, nous utilisons des valeurs booléennes afin de déterminer la présence d'un lien ou non. Deux utilisateurs similaires sont ainsi reliés par un arc dont le poids vaut 1, sinon ce poids vaut 0 (i.e. le lien est inexistant). Autrement dit, le réseau comportemental est représenté par un graphe non valué. Après le calcul des plus courts chemins, nous déduisons les valeurs de similarités en utilisant l'équation (1.7). Ici, d est représenté par le nombre de nœuds intermédiaires et d_{max} désigne le nombre maximal de nœuds reliant un utilisateur avec l'un de ses voisins.

L'application des méthodes de prédiction de lien mène à l'identification de nouveaux liens entre utilisateurs, au niveau du réseau comportemental. La plupart des méthodes présentées ci-dessus (à l'exception de *l'attachement préférentiel*) applique le principe de transitivité lors de l'exploration des nouveaux liens.

La figure 1.4 présente un exemple d'identification de voisins par le D-BNCF grâce à l'exploitation des méthodes de prédiction de lien. Pour des raisons de simplification, les valeurs numériques n'ont pas été fournies dans les matrices.
Sur cette figure, si nous comparons la matrice du BNCF aux matrices D-BNCF, nous observons que l'application des méthodes de prédiction de lien mène à une augmentation du nombre de voisins. Notons que les matrices relatives à D-BNCF-AttachementPréférentiel et D-BNCF-DistanceDeGraphe sont pleines. Ce résultat découle du fait qu'un lien entre deux utilisateurs est créé lorsqu'un utilisateur a au moins un voisin.

FIG. 1.4 – Exemple comparant les voisins identifiés par D-BNCF (selon les méthodes de prédiction de lien)

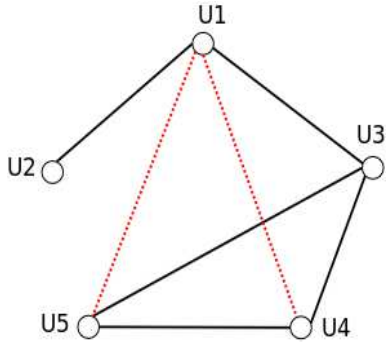


En comparant les matrices D-BNCF, nous observons que dans certains cas, des liens sont découverts par toutes les méthodes, tel que le lien entre u_1 et u_4. Dans d'autres cas, selon la méthode utilisée, les nouveaux liens peuvent être rajoutés ou pas. En effet, si nous comparons les matrices D-BNCF-Adamic/Adar et D-BNCF-Jaccard, nous remarquons que D-BNCF-Jaccard a identifié un nouveau lien entre u_2 et u_3, alors que D-BNCF-Adamic/Adar n'a pas identifié ce lien. En effet, le voisin commun u_1 de la paire de nœuds $\langle u_2, u_3 \rangle$ n'est pas fréquent parmi les autres paires de nœuds.

La figure 1.5 présente la matrice D-BNCF-Adamic/Adar sous forme de graphe. Les lignes pleines représentent les liens directs calculés par le BNCF et les lignes pointillées

représentent les nouveaux liens découverts par le D-BNCF-Adamic/Adar.
Nous observons d'abord que l'utilisateur u_1 a deux voisins directs u_2 et u_3. L'application de la méthode Adamic/Adar a contribué à l'augmentation du voisinage par deux nouveaux voisins u_4 et u_5.
Ainsi, bien que ces deux utilisateurs ne soient pas similaires en terme de navigation ou de comportement (i.e. ils n'ont pas consulté suffisamment d'items en commun dans le passé), un lien potentiellement fort entre eux est susceptible d'être découvert grâce à la forte similarité de leurs voisins intermédiaires.

FIG. 1.5 – Identification de nouveaux voisins par D-BNCF-Adamic/Adar



Les différentes méthodes présentées dans cette section permettent d'identifier de nouveaux voisins potentiels pour tous les utilisateurs actifs. Ces voisins sont par la suite impliqués dans le calcul des prédictions dans l'objectif de remédier au manque de données et d'améliorer la performance du système de recommandation.

1.2.3 Génération des prédictions

Une fois les nouveaux liens entre un utilisateur actif u_a et les autres utilisateurs sont identifiés (PHASE I), les prédictions sont calculées (PHASE II) en se basant sur l'équation de la somme pondérée utilisée dans les chapitres précédents, afin de calculer les prédictions pour chaque utilisateur actif.
Nous sélectionnons les plus proches voisins U_a (directs et non directs) dans le réseau comportemental, ayant déjà noté l'item à prédire i_k.
Les items qui seront recommandés à l'utilisateur actif sont les items disposant des valeurs de prédictions les plus élevées.

1.3 Evaluation du modèle

En vue d'évaluer la performance du D-BNCF, nous avons utilisé le même corpus d'usage du Crédit Agricole expérimenté dans les chapitres précédents. Ce corpus inclut

les traces de navigation des utilisateurs du Groupe Crédit Agricole. Dans le cadre de cette évaluation, nous avons utilisé les métriques MAE et HMAE afin d'évaluer la précision des recommandations.

1.3.1 Modèles expérimentés

L'objectif de cette évaluation consiste à étudier l'impact de chaque méthode de prédiction de lien sur la performance du système de recommandation.
Les modèles D-BNCF sont notés ainsi :

- D-BNCF-VoisinsCommuns,
- D-BNCF-AttachementPréférentiel,
- D-BNCF-Jaccard,
- D-BNCF-Adamic/Adar,
- D-BNCF-ETL(Enhanced Transtitive Link),
- D-BNCF-DistanceDeGraphe-Valué (en considérant les similarités en tant que poids des arcs),
- D-BNCF-DistanceDeGraphe-NonValué (les poids des arcs ne sont pas considérés, c'est le nombre de nœuds séparant deux utilisateurs qui détermine le plus court chemin).

En outre, nous avons comparé la performance des modèles D-BNCF au :

- BNCF : il représente le réseau comportemental sans aucun nouveau lien. Seuls les voisins directs sont impliqués au calcul des prédictions.
- FCS.

Par ailleurs, dans l'objectif d'examiner si les modèles D-BNCF peuvent être complémentaires, nous avons proposé un autre modèle combinant les similarités calculées par les deux modèles D-BNCF les plus performants.

1.3.2 Résultats

MAE

Le tableau 1.1 présente les résultats en MAE relatifs aux modèles D-BNCF comparés au BNCF ainsi qu'au FCS. En observant les résultats du tableau 1.1, nous remarquons qu'en cas d'application de la méthode distance de graphe (D-BNCF-DistanceDeGraphe-Valué/D-BNCF-DistanceDeGraphe-NonValué) exploitant soit un graphe (réseau comportemental) valué ou bien non valué, l'utilisation de nouveaux liens contribue à une légère

amélioration, comparée au BNCF exploitant uniquement les liens directs.
Or, la précision en MAE se détériore lorsque les méthodes d'Attachement préférentiel, de voisins communs, de Jaccard, d'Adamic/Adar et d'ETL (Enhanced Transtitive Link) sont appliquées au réseau comportemental.

Tab. 1.1 – Résultats en MAE

Modèles de recommandation	**MAE**
FCS	**0.763**
BNCF	**0.789**
D-BNCF-VoisinsCommuns	1.074
D-BNCF-AttachementPréférentiel	1.011
D-BNCF-Jaccard	0.858
D-BNCF-Adamic/Adar	0.882
D-BNCF-ETL	0.847
D-BNCF-DistanceDeGraphe-Valué	**0.782**
D-BNCF-DistanceDeGraphe-NonValué	**0.780**

En outre, nous constatons qu'avec l'application des méthodes Jaccard, Adamic/Adar, ETL et distance de graphe, nous obtenons une meilleure précision que les méthodes Common neighbors et Attachement préférentiel. La performance de ces méthodes résulte de la technique utilisée pour la découverte de nouveaux liens. En effet, l'Attachement préférentiel a pour limite de créer un réseau hyper-connecté. Par conséquent, de nombreux voisins (même ceux qui ne sont pas réellement similaires aux utilisateurs actifs) sont intégrés au calcul des prédictions.
En ce qui concerne la méthode voisins communs, elle a comme inconvénient d'attribuer des valeurs de similarités élevées entre les utilisateurs ayant beaucoup de voisins, au détriment de ceux ayant un nombre faible de voisins.

De plus, en cas d'application du D-BNCF-Jaccard, les utilisateurs sont considérés comme similaires lorsqu'ils partagent une importante proportion de voisins communs parmi tous leurs voisins. Cette méthode s'avère ainsi plus performante, parmi les autres méthodes citées plus haut. La méthode ETL, qui constitue une amélioration de la méthode Jaccard (en considérant spécifiquement les voisins communs les plus proches), mène à une légère amélioration de la qualité des recommandations, comparée à Jaccard, en terme de MAE.
En ce qui concerne le D-BNCF-Adamic/Adar, sa performance est liée particulièrement au fait que les voisins communs rares ont un poids plus important que les voisins fréquents. Si deux utilisateurs ont en commun des voisins rares, ils tendent à être très similaires.

HMAE

Les résultats en HMAE relatifs à cette expérimentation sont présentés dans le tableau 1.2.

Si nous observons les résultats des modèles D-BNCF, nous pouvons signaler d'abord que l'application des méthodes de prédiction de lien contribue à une importante amélioration de la précision des recommandations en terme de HMAE.
Comparé au FCS, l'utilisation des méthodes de prédiction de lien améliore la HMAE d'environ 33%. De plus, l'utilisation des méthodes Jaccard, Adamic/Adar ou distance de graphe (valué et non valué), mène à une meilleure précision de 24%, 27% et 7% respectivement, comparé au BNCF. Cependant, l'utilisation de l'Attachement préférentiel et de l'ETL diminue faiblement la précision du BNCF en terme de HMAE. La méthode de voisins communs, quant à elle, bien qu'elle engendre un taux d'erreur plus élevé au niveau des recommandations comparée aux autres méthodes, reste tout de même plus performante que le modèle FCS.

Notons que Jaccard, Adamic/Adar et distance de graphe contribuent à de meilleures performances en HMAE, comparés aux méthodes d'attachement préférentiel et de voisins communs. Ces performances confirment la fiabilité des méthodes Jaccard, Adamic/Adar et distance de graphe pour l'évaluation des similarités entre les paires d'utilisateurs à travers le réseau comportemental.

Tab. 1.2 – Résultats en HMAE

Modèles de recommandation	**HMAE**
FCS	0.541
BNCF	0.501
D-BNCF-VoisinsCommuns	0.536
D-BNCF-AttachementPréférentiel	0.505
D-BNCF-Jaccard	**0.380**
D-BNCF-Adamic/Adar	**0.364**
D-BNCF-ETL	0.515
D-BNCF-DistanceDeGraphe-Valué	0.468
D-BNCF-DistanceDeGraphe-NonValué	0.471

Dans le but d'évaluer la complémentarité entre des modèles D-BNCF en terme de prédiction et en prenant en considération les résultats des modèles D-BNCF présentés ici, nous avons choisi d'évaluer en outre, un autre modèle combinant les similarités issues de deux modèles D-BNCF. Ces deux modèles exploitent différentes méthodes de prédiction de lien. Les résultats de cette évaluation sont présentés dans ce qui suit.

1.3.3 D-BNCF Combiné

En tenant compte des performances des modèles D-BNCF décrites ci-dessus, nous avons sélectionné les modèles D-BNCF exploitant les méthodes Jaccard et Adamic/Adar, vu leur performance en terme de HMAE.
Bien que Jaccard et Adamic/Adar appartiennent à la même famille de méthodes de prédiction de lien, à savoir les méthodes fondées sur le voisinage des nœuds, ces méthodes

mesurent différemment les liens entre utilisateurs. Jaccard considère le critère des voisins communs et Adamic/Adar prend en compte les voisins communs rares. A cet effet, la combinaison des similarités provenant de ces modèles reste cohérente et potentiellement complémentaire. La combinaison des similarités résultant de ces deux méthodes est ainsi susceptible d'améliorer la qualité des recommandations.
Notons que si une paire de nœuds u_a et u_b est corrélée au niveau des deux modèles D-BNCF-Jaccard et D-BNCF-Adamic/Adar, nous retenons la moyenne des valeurs de similarité calculées par chacun de ces deux modèles.

Les résultats relatifs à l'expérimentation du D-BNCF combiné correspondent à 0.870 en terme de MAE et à 0.355 en HMAE. Ainsi, au niveau de la MAE, aucune amélioration de la précision n'est atteinte. Or, en terme de HMAE, le modèle combiné a permis d'atteindre une meilleure précision comparée à tous les modèles étudiés ici. Ce modèle améliore la précision de 3%, comparé au meilleur score de précision obtenu auparavant parmi tous les modèles. De plus, au niveau de la HMAE, le modèle combiné contribue à une amélioration de 34% comparé au FCS et de 29% comparé au BNCF. Ces résultats confirment la complémentarité entre les méthodes Jaccard et Adamic/Adar et la pertinence de leur combinaison.

1.3.4 Discussion

Nous avons présenté ici le modèle de recommandation D-BNCF que nous avons proposé. Le D-BNCF exploite un réseau comportemental (construit à partir des similarités de comportement entre utilisateurs) ainsi que les méthodes de prédiction de lien permettant de densifier ce réseau. L'objectif du D-BNCF consiste à découvrir de nouveaux liens entre utilisateurs. Ces nouveaux liens sont impliqués dans le processus de recommandation afin de pallier le manque de données et d'améliorer la qualité des recommandations.

L'évaluation des modèles D-BNCF montre l'impact des méthodes de prédiction de lien Jaccard et Adamic/Adar, en particulier en terme de HMAE. La performance de ces deux méthodes est liée à la façon dont les nouveaux liens sont identifiés, considérant les voisins communs et les voisins communs rares, plutôt que plusieurs voisins intermédiaires telle que dans la méthode distance de graphe.
La faible précision des recommandations produites par le D-BNCF-AttachementPréférentiel et D-BNCF-VoisinsCommuns était prévisible. En effet, ces deux méthodes engendrent respectivement une hyperconnectivité du réseau comportemental ainsi que l'augmentation de l'impact des utilisateurs disposant de nombreux voisins.

En ce qui concerne la méthode de prédiction que nous avons proposée ETL (Enhanced Transitive Link), la performance réalisée reste modeste. Les résultats obtenus au niveau de la précision des recommandations sont notamment dûs à la stratégie de sélection des plus proches voisins communs.
En vue d'examiner cette question et dans la perspective d'améliorer la précision des re-

commandations, nous avons réalisé une autre expérimentation du D-BNCF-ETL, en considérant une autre stratégie pour la découverte de nouveaux liens entre utilisateurs. Cette stratégie consiste à sélectionner les TopN voisins communs les plus proches à partir de la valeur médiane de similarité, calculée pour le voisinage de chaque utilisateur. Nous avons considéré ces TopN, tel que $N = 60\%$, $N = 40\%$ ou $N = 20\%$ des meilleurs voisins.
Les meilleurs résultats relatifs à cette expérimentation correspondent au choix du $TopN40\%$ pour la sélection des plus proches voisins communs. Le résultat obtenu pour cette expérimentation correspond à 0.849 en terme de MAE et à 0.347 en HMAE. Ainsi, en prenant en compte les autres expérimentations relatives aux modèles D-BNCF, nous pouvons déduire que la performance de la méthode ETL est très sensible aux stratégies de sélection des voisins communs les plus proches.

En outre, au niveau de la méthode distance de graphe, nous avons montré que la considération du nombre de nœuds, séparant deux utilisateurs sur le chemin le plus court peut être aussi fiable que la considération des poids de similarités comportementales. En effet, les résultats de l'expérimentation soulignent l'importance de cette méthode pour prédire efficacement les liens entre utilisateurs.

De plus, l'évaluation du D-BNCF combiné, présentée ci-dessus, met en évidence l'importance de combiner les similarités calculées par les modèles D-BNCF-Jaccard et D-BNCF-Adamic/Adar. Cette combinaison s'avère en effet complémentaire, considérant que ces deux modèles exploitent deux méthodes différentes pour évaluer les similarités entre utilisateurs.

Par ailleurs, comme pour les réseaux sociaux, les réseaux comportementaux sont dynamiques et ont tendance à évoluer rapidement par l'ajout de nouveaux nœuds. Ainsi, pour résoudre la question d'évolution des réseaux comportementaux, une stratégie différente (que celle présentée ici) doit être appliquée pour l'exploitation des méthodes de prédiction de lien. En effet, le processus de découverte de nouveaux liens dans le cadre de larges réseaux, peut être limité par exemple à un sous-ensemble de nœuds (utilisateurs), tels que les nœuds ne disposant pas d'un nombre suffisant de voisins.

En outre, une question qui reste également à résoudre est le démarrage à froid. En effet, en cas d'introduction d'un nouvel item au système, cet item ne disposant pas encore d'appréciations de la part des utilisateurs, ne peut être intégré dans le processus de recommandation. Dans cette perspective, nous avons proposé un modèle qui repose sur les leaders comportementaux pour la recommandation de la nouveauté. Ce modèle est décrit dans le chapitre suivant.

Chapitre 2

Leaders comportementaux pour la recommandation de la nouveauté

Dans les chapitres précédents, nous nous sommes intéressés à l'étude des problèmes de manque de données et de la qualité des recommandations. Dans le cadre des systèmes de recommandation fondés sur le FCS, une autre question de recherche qui demeure soulevée est le démarrage à froid concernant les items, appellée aussi problème de latence [Sollenborn et Funk, 2002]. En effet, un item récemment intégré à un système de recommandation, n'étant pas encore consulté ou noté par un utilisateur, ne peut être recommandé aux utilisateurs actifs.
En vue de résoudre ce problème de latence, la solution la plus communément utilisée consiste à exploiter la technique basée sur le contenu (cf. section 1.4.2, chapitre 1, partie 1). Lorsqu'un nouvel item est intégré, le système évalue sa similarité avec les autres items disponibles en terme de contenu. Ainsi, ce nouvel item pourra être recommandé à un utilisateur ayant apprécié dans le passé des items ayant un contenu similaire à ce nouvel item.
La technique basée sur le contenu constitue un moyen d'amorçage et permet de recommander un nouvel item dès son intégration dans le système. Or, sur le long terme, l'utilisation de cette technique peut ne pas être appropriée. En effet, la technique basée sur le contenu a pour inconvénient d'engendrer une surspécialisation des recommandations (i.e. toutes les recommandations sont liées à un même domaine). De plus, cette technique pose des problèmes lorsqu'il s'agit d'items qui ne sont pas des données textuelles.

Dans ce chapitre, nous présentons le modèle que nous avons proposé dans le but d'atténuer ou de réduire le temps de latence. Ce modèle repose sur l'identification de leaders comportementaux dans le contexte des réseaux comportementaux [Esslimani *et al.*, 2010b] et des systèmes de recommandation [Esslimani *et al.*, 2010c].

Dans le domaine des réseaux sociaux, un leader est une personne qui influence ses amis ou ses collaborateurs par ses idées et ses opinions. Ici, nous considérons qu'un leader

comportemental est un utilisateur fortement connecté à des utilisateurs ayant un comportement similaire et qui prédit fiablement les appréciations de ces utilisateurs. A notre sens, en connaissant leurs opinions sur les nouveaux items, ces leaders représentent les utilisateurs qu'un système de recommandation doit cibler afin de prédire les appréciations des autres utilisateurs du réseau concernant ces items.

Dans les sections qui suivent, nous présenterons d'abord quelques travaux de recherche ayant trait au leadership et à la détection de leaders et d'influenceurs. Par la suite, nous décrirons l'algorithme proposé pour la détection de leaders comportementaux ainsi que les résultats de son évaluation.

2.1 Détection des leaders et des influenceurs

Le leadership et la propagation de l'influence ont fait l'objet de nombreuses études liées au domaine du marketing, des sciences sociales et de l'analyse des réseaux sociaux [Goyal *et al.*, 2008]. Ces études visent à comprendre comment les communautés émergent, quelles sont leurs propriétés, comment elles évoluent, quels sont les rôles des membres de ces communautés et comment les influenceurs ou les leaders d'opinion peuvent être détectés à travers ces communautés.
Katz et Lazarsfeld [Katz et Lazarsfeld, 1955] ont défini les leaders d'opinion comme "les individus qui sont susceptibles d'influencer les autres personnes appartenant à leur environnement immédiat". Les premières études de l'influence et du leadership ont mis l'accent sur l'analyse de la propagation des innovations médicales et technologiques [Coleman *et al.*, 1966]. Plus récemment, [Valente, 1995] a examiné également cette question en proposant des modèles de diffusion de l'innovation dans le cadre de réseaux.

Dans le domaine du marketing (marketing viral), la propagation de l'influence est souvent liée au phénomène du "bouche-à-oreille" et à son effet sur le succès de nouveaux produits [Domingos et Richardson, 2001].
Le challenge le plus important en marketing est comment trouver un petit segment de la population (influenceurs ou leaders) capable d'influencer les autres segments, par leurs opinions positives ou négatives concernant des produits ou des services [Watts et Dodds, 2007]. Keller et Berry [Keller et Berry, 2003] confirment l'importance des influenceurs dans la mesure où ils orientent les décisions d'une communauté et prédisent les futures tendances de marchés. Selon leur étude, "un américain sur dix dit aux neuf autres comment voter, où manger et quoi acheter".

Avec le développement de l'Internet, les leaders et les influenceurs n'utilisent pas uniquement le bouche-à-oreille traditionnel, ils peuvent propager leurs opinions à travers des échanges interactifs sur les blogs, les forums, les wikis et les différentes plate-formes de réseaux sociaux. En effet, de nos jours, les réseaux sociaux deviennent le media le plus important pour la propagation d'informations, d'innovations et d'opinions.

De nombreuses études récentes se sont intéressées à l'analyse des interactions et des influences entre entités et à l'évaluation de l'impact des leaders dans les réseaux sociaux. Par exemple, [Kempe *et al.*, 2003] ont étudié les algorithmes d'approximation pour la maximisation d'influence dans les réseaux de co-auteurs. [Agarwal *et al.*, 2008] s'intéressent à l'identification des blogueurs influenceurs actifs et non actifs permettant d'orienter les tendances et d'affecter les intérêts de groupes dans le contexte des blogs. [Goyal *et al.*, 2008] proposent une approche d'analyse de motifs afin de découvrir les leaders et d'évaluer leur influence sur le réseau social. Des actions telles que le tagging, l'attribution de note, l'achat ou l'envoi d'un message sur un blog, sont considérées lors de la découverte des motifs fréquents. [Goyal *et al.*, 2008] considèrent en effet que dans un réseau social, un leader peut guider les tendances de réalisation d'actions. Ainsi, les amis sont tentés de réaliser les mêmes actions que celles effectuées par le leader.

Par ailleurs, d'autres études ont été dédiées à l'étude de l'impact de la structure de réseau sur la propagation d'informations et d'opinions. [Barabási *et al.*, 2002] [Newman, 2003] mettent notamment en évidence le rôle des nœuds hyperconnectés dans un réseau social (appelés également hubs), pour la diffusion d'information et pour l'évolution de la collaboration dans ce réseau. [Gladwell, 2000] confirme également que les nœuds très connectés ont une influence considérable sur leurs voisins. Keller et Berry [Keller et Berry, 2003] montrent aussi que les utilisateurs ayant une influence sur les autres, disposent relativement d'un nombre élevé de liens sociaux.

A notre connaissance, dans le contexte des systèmes de recommandation et du FCS, la détection de leaders a été examinée dans peu d'études. Parmi ces études, nous pouvons citer le travail de [Cheon et Lee, 2005], dont l'objectif consiste à résoudre le problème de démarrage à froid lié à un nouvel utilisateur. [Cheon et Lee, 2005] proposent ainsi un système de recommandation permettant de sélectionner les leaders d'opinion. Afin de détecter ces leaders, ce système utilise des inférences exploitant une méthode issue du marketing nommée RFM (Recency, Frequency, Monetary). Par la suite, les topN items appréciés par les leaders identifiés sont proposés à un nouvel utilisateur.
[O'Reilly, 2005] a défini différentes métriques permettant de mesurer l'influence des utilisateurs sur les systèmes de recommandation exploitant les notes. Ils proposent notamment une métrique qui mesure l'influence en supprimant les notes de certains utilisateurs lors du calcul des prédictions afin d'observer l'effet de cette suppression sur les résultats des recommandations. Si la différence est importante, l'utilisateur est détecté comme étant influenceur.

Ce qui distingue le modèle proposé ici des travaux cités ci-dessus est que la détection de leaders n'est pas une fin en soi, mais elle va nous servir à atténuer le problème de latence ou de nouveauté des items. De plus, afin de déterminer les leaders les plus fiables, notre approche repose sur deux critères. Le premier critère est lié aux approches utilisées dans le domaine de l'analyse des réseaux sociaux. Le deuxième critère consiste à analyser la capacité prédictive d'un utilisateur ou d'un leader potentiel en exploitant les liens comportementaux. Notre approche va être décrite dans la section suivante.

2.2 Détection des leaders comportementaux

Les systèmes de recommandation fondés sur le Filtrage Collaboratif Standard (FCS) requièrent un volume considérable de données de notes afin d'évaluer les similarités entre utilisateurs et calculer les recommandations. Lorsqu'un item est nouveau, les notes relatives à cet item ne sont pas encore disponibles. Par conséquent, le système ne peut incorporer cet item dans les listes de recommandation. De plus, si ce nouvel item est peu noté par les utilisateurs, il y a peu de chance à ce qu'il soit recommandé.
Dans le but de réduire le temps de latence, nous proposons d'identifier les leaders dans le cadre d'un réseau comportemental. Les appréciations de ces leaders sont par la suite propagées au travers de ce réseau comportemental en vue de prédire les avis des autres utilisateurs sur les nouveaux items et éventuellement leur recommander ces items.

Contrairement au FCS, notre système de recommandation ne nécessite pas plusieurs notes concernant les nouveaux items afin de les incorporer parmi les recommandations. Seule l'information parvenant des leaders concernant ces items suffit.
En outre, dans les réseaux sociaux, la détection de leaders repose sur l'analyse des liens sociaux à travers le réseau. Ici, nous considérons l'analyse des liens comportementaux dans l'objectif d'identifier les leaders. Pour la construction du réseau comportemental, nous avons utilisé la même modélisation que celle décrite précédemment.

En s'appuyant notamment sur les études de [Gladwell, 2000], [Barabási *et al.*, 2002], [Newman, 2003] et [Keller et Berry, 2003] mentionnées précédemment, nous définissons un leader comportemental comme étant un utilisateur, qui n'est pas seulement hyperconnecté dans le réseau comportemental, mais qui dispose également d'un important potentiel de prédiction des futures appréciations des autres utilisateurs.
Nous supposons en effet qu'un leader comportemental peut propager ses appréciations dans le réseau. Nous proposons de propager ces appréciations en utilisant un facteur d'atténuation. Ce facteur est lié directement à la similarité entre utilisateurs (les poids des liens). En effet, quand les utilisateurs sont très similaires, nous considérons qu'il existe une grande probabilité qu'ils aient des appréciations semblables concernant les items.

En outre, dans une approche classique de FC les items recommandés par les systèmes de recommandation aux utilisateurs actifs sont les items appréciés par leurs voisins. Ainsi, de la même façon, nous supposons qu'un leader comportemental peut propager (recommander) les items qu'il apprécie. De ce fait, il a été nécessaire de décomposer l'ensemble des appréciations des utilisateurs (leaders potentiels) selon une échelle binaire : "aime" ou "aime pas" un item. Selon l'échelle $[1-5]$ par exemple, nous considérons que les notes 4 et 5 correspondent aux items appréciés, tandis que les notes 1, 2 et 3 correspondent aux items non appréciés.

L'algorithme 4 représente l'algorithme que nous proposons pour la détection des leaders comportementaux. Cet algorithme utilise en entrée le graphe modélisant le réseau comportemental, où les nœuds représentent les utilisateurs et les arcs sont les liens les

reliant. Notre algorithme inclut deux étapes majeures. Dans chaque étape, des sous-ensembles distincts d'items notés I_{tr} et I_{ts} sont considérés. I_{tr} correspond aux items utilisés (dans la phase d'apprentissage) pour évaluer les similarités de comportement et pour construire le réseau comportemental. I_{ts} représente le sous-ensemble des nouveaux items exploités pour la validation des leaders comportementaux effectifs (la phase de test).

Algorithm 4 Détection de leaders comportementaux

function SELECTIONNERLEADERSPOTENTIELS
  for chaque nœud u_a dans le graphe G **do**
    Evaluer "Degré de centralité " $D_{(u_a)}$ ▷ noté $|\Gamma_{(u_a)}|$

$$D_{(u_a)} = |\Gamma_{(u_a)}| \tag{2.1}$$

  end for
  Trier les degrés D de tous les nœuds N dans un ordre descendant
  return TopN leaders potentiels U_{PL} ayant un degré de centralité élevé
end function

function DETECTERLEADERS
  for chaque leader potentiel $u_{pl} \in U_{PL}$ **do**
    Sélectionner les items appréciés $I_{prf}(u_{pl}) \subset I_{ts}$
    Sélectionner les nœuds voisins
    for chaque voisin sélectionné u_a **do**
      for chaque item $i_j \in I_{prf}(u_{pl})$ **do**
        Propager les appréciations $apr(u_{pl}, i_j)$ à u_a tel que :

$$papr(u_a, i_j) = \alpha_{(u_a, u_{pl})} * apr(u_{pl}, i_j) \tag{2.2}$$

        Evaluer la précision de chaque $papr(u_a, i_j)$ ▷ $papr(u_a, i_j)$ est pertinent ou non pour u_a
      end for
      Evaluer la précision de toutes les appréciations propagées à u_a
    end for
    Evaluer la précision du leader potentiel u_{pl} comme la moyenne des précisions p calculées parmi tous ses voisins

$$P(u_{pl}) = \frac{\sum_{u_a=1}^{m} p}{m} \tag{2.3}$$

  end for
  return TopN leaders comportementaux effectifs U_L ayant les meilleurs ratios de précision
end function

Dans la première étape de l'algorithme (fonction "SélectionnerLeadersPotentiels"), pour chaque nœud u_a dans le graphe, la connectivité ou le degré de centralité est calculé comme étant le nombre de liens (voisins) incidents à u_a. Par la suite, les TopN leaders

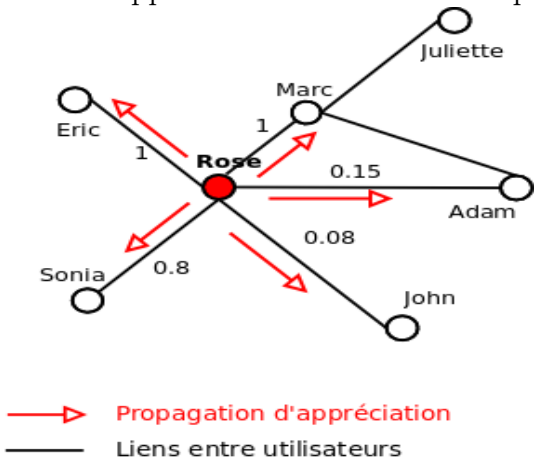
potentiels U_{PL} sont sélectionnés en prenant en considération leur forte connectivité dans le réseau comportemental.

Dans la deuxième étape de l'algorithme (fonction "DétecterLeaders"), pour chaque leader potentiel $u_{pl} \in U_{PL}$, les items appréciés sont identifiés $I_{prf}(u_{pl}) \subset I_{ts}$. Par la suite, selon l'équation (2.2), les appréciations du leader potentiel $apr(u_{pl}, i_j)$ concernant les items i_j ($i_j \in I_{prf}(u_{pl})$) sont propagées aux voisins directs tel qu'une appréciation propagée, notée $papr(u_{pl}, i_j)$, d'un leader u_{pl} à un nœud voisin u_a concernant l'item i_j, est pondérée par le coefficient $\alpha_{(u_a,u_{pl})}$. Les poids α varient de 0 à 1 selon la similarité entre u_{pl} et u_a.

Une fois les appréciations propagées à un voisin u_a, elles sont évaluées en terme de précision (cf. section 2.4.2, chapitre 2, partie 1). Par la suite, pour chaque leader potentiel, nous évaluons la précision $P(u_{pl})$ en utilisant l'équation (2.3). Cette précision est équivalente à la moyenne des précisions calculées parmi tous ses voisins u_a. Notons que m désigne, dans l'équation (2.3), le nombre de voisins de u_{pl}.
Ainsi, les ratios de précision obtenus permettent de mettre en évidence les leaders comportementaux effectifs. Plus le ratio de précision est élevé, plus le leader est fiable.

Afin d'illustrer le processus de propagation, nous présentons l'exemple de la figure 2.1 qui représente la propagation d'appréciation concernant des "articles d'actualité". Considérant son importante connectivité dans le réseau comportemental ($D_{(Rose)} = 5$), Rose est un leader comportemental potentiel parmi les autres utilisateurs. Lorsque Rose propage son appréciation sur l'article "Web 2.0 applications", les valeurs de similarité de Rose avec les autres utilisateurs sont considérées. A cet effet, Marc, Eric et Sonia vont recevoir une recommandation de cet article, puisque $SimNav(Rose, Eric) = 1.0$, $SimNav(Rose, Marc) = 1.0$ et $SimNav(Rose, Sonia) = 0.8$. Toutefois, Adam et John reçoivent une appréciation négative concernant le même article, au vu de leur faible similarité avec Rose ($SimNav(Rose, Adam) = 0.15$ et $SimNav(Rose, John) = 0.08$). Ainsi, l'article "Web 2.0 applications" ne sera pas recommandé aux utilisateurs Adam et John.

FIG. 2.1 – Propagation de l'appréciation d'un leader comportemental potentiel

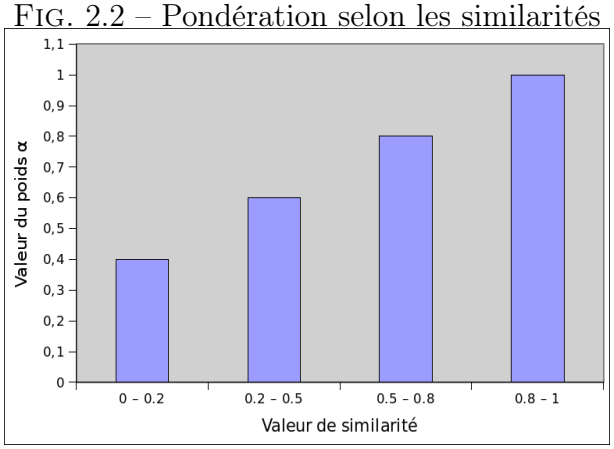


De ce fait, lorsque le système de recommandation a besoin de générer des recommandations concernant les nouveaux items, les leaders comportementaux détectés par notre algorithme sont considérés. En effet, comme ces leaders représentent les nœuds ou les points d'entrée dans le réseau comportemental, le système de recommandation recommande ces nouveaux items à ces leaders. Ainsi, si ces leaders attribuent des avis positifs quant aux nouveaux items, ils font un "push" de leurs appréciations à leurs voisins en utilisant l'équation (2.2).

2.3 Evaluation des recommandations de leaders

Dans l'objectif d'évaluer la performance de l'approche présentée ici, nous avons exploité le corpus d'usage du Crédit Agricole qui a été également utilisé dans les expérimentations précédentes.
Afin de valider la qualité des appréciations propagées par les leaders potentiels à travers le réseau, nous avons extrait ces appréciations du corpus test nommé I_{ts}. Comme nous l'avons mentionné auparavant, nous considérons uniquement les appréciations positives de ces leaders (seuls les items qu'ils apprécient, notés 4 et 5).
De plus, les poids α sont utilisés dans l'étape de propagation comme un facteur d'atténuation. Ces poids varient de 0 à 1. A titre d'exemple, lorsque les valeurs de similarités appartiennent à l'intervalle $]0.8 - 1.0]$, le poids α correspondant vaut 1.0. Notons que l'attribution des poids α, présentée dans la figure 2.2, repose sur la distribution des similarités entre utilisateurs relative au corpus étudié ici.

FIG. 2.2 – Pondération selon les similarités



2.3.1 Résultats

Dans cette expérimentation, nous avons évalué la précision des appréciations propagées de chaque leader potentiel en utilisant l'équation (2.3).

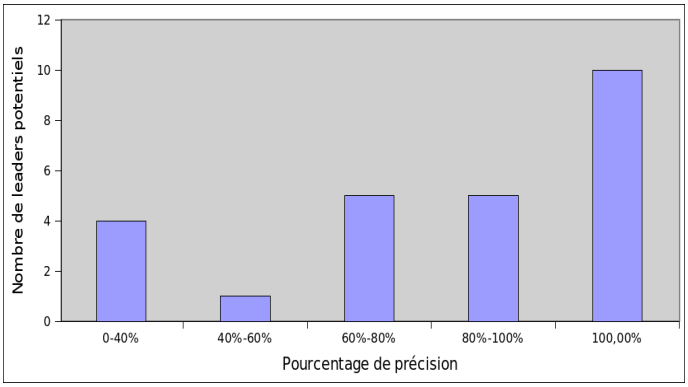
Les figures 2.3 et 2.4 présentent les distributions du nombre de leaders comportementaux potentiels en fonction de la précision, en prenant en considération respectivement 10% et 20% des TopN leaders potentiels lors de la propagation.
Les TopN10 et TopN20 correspondent respectivement à 53 et à 101 leaders potentiels parmi tous les utilisateurs dans le corpus étudié (748 utilisateurs). Le choix des TopN10 et TopN20 est lié au fait que notre objectif est d'identifier un petit segment d'utilisateurs (leaders) capables de prédire efficacement les appréciations des voisins.
Notons que pour environ 53% des TopN10 leaders et 49% des TopN20 leaders, la précision ne peut être évaluée au vu des raisons suivantes :

- Les items recommandés par les leaders comportementaux potentiels n'ont pas été encore consultés par leurs voisins. Ainsi, nous ne pouvons pas déterminer si les leaders potentiels sont fiables ou non.

- Les leaders comportementaux potentiels ne disposent pas d'appréciations positives (dans le corpus test I_{ts}). Par conséquent, ils ne peuvent pas effectuer de propagation envers leurs voisins.

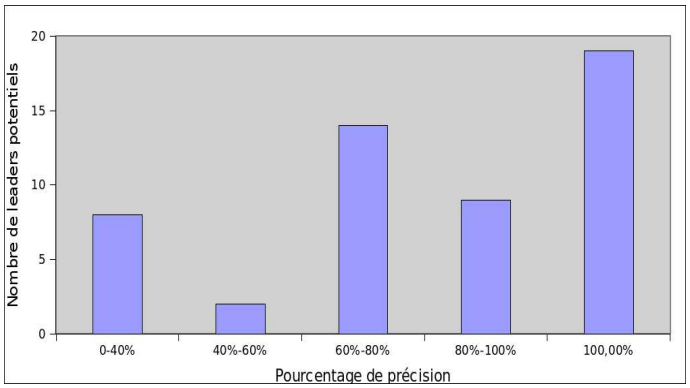
Il est à signaler que dans les résultats présentés ici, cette catégorie de leaders n'est pas considérée.

FIG. 2.3 – Distribution des TopN10 leaders comportementaux potentiels selon le pourcentage de précision



Si nous observons les résultats des figures 2.3 et 2.4, nous remarquons que les distributions de précision ont une évolution similaire pour les TopN10 et TopN20 leaders potentiels. Lorsque les TopN10 leaders comportementaux sont impliqués, nous observons que 80% de ces leaders ont plus de 60% de précision, 60% ont une précision de plus de 80% et 40% ont atteint 100% de précision.
En ce qui concerne les TopN20 leaders comportementaux, nous observons que, de la même façon, environ 80% de leaders propagent efficacement les recommandations, puisque la précision correspondante est supérieure à 60%, 53% ont une précision supérieure à 80% et 37% ont une précision qui s'élève à 100%.

FIG. 2.4 – Distribution des TopN20 leaders comportementaux potentiels selon le pourcentage de précision



Avec l'utilisation des TopN10 ou des TopN20, une importante proportion de leaders comportementaux potentiels obtient une grande précision relative aux appréciations propagées. Nous considérons que les leaders ayant atteint plus de 80% de précision, constituent les nœuds représentatifs parmi tous les nœuds dans le réseau comportemental. En effet, ils prédisent efficacement les appréciations des autres utilisateurs.

En outre, dans cette expérimentation nous avons comparé la performance de notre modèle à la performance du FCS (Filtrage Collaboratif Standard), en terme de précision (cf. section 2.4.2, chapitre 2, partie 1). Le tableau 2.1 présente les moyennes de précision correspondant à notre modèle "Recommandations fondées sur les leaders" ainsi qu'au FCS. Ces précisions ont été calculées sur les mêmes paires $< utilisateur, item >$ en utilisant deux ensembles différents R_1 et R_2. Ces ensembles représentent respectivement les paires prédites $< utilisateur, item >$, considérées lors de la propagation par les TopN10 et les TopN20 leaders.
En observant les résultats du tableau 2.1, nous remarquons qu'au niveau des items recommandés par les leaders (contenus dans R_1 et R_2), notre modèle mène à une meilleure performance comparé au FCS. En effet, lorsque nous considérons les ensembles R_1 et R_2, environ 77% de précision est atteinte. Cependant, le FCS est moins performant puisqu'il parvient uniquement à 51% et à 43% de précision, en considérant respectivement R_1 et R_2. Ces résultats confirment ainsi la fiabilité des leaders comportementaux pour la recommandation d'items pertinents aux autres utilisateurs.

TAB. 2.1 – Moyenne de précision des recommandations fondées sur les leaders comparée au FCS

Modèle de recommandation	R_1	R_2
Recommandations fondées sur les leaders	77%	76%
FC Standard (FCS)	51%	43%

2.3.2 Discussion

Dans ce chapitre, nous avons présenté le modèle de recommandation qui a été proposé en vue de réduire le temps de latence. Habituellement, les travaux de recherche traitant du problème de latence exploitent la technique basée sur le contenu. Or, la considération uniquement du contenu des items dans le cadre des recommandations présente quelques limites, dont le manque de diversité des recommandations générées par le système.
Notre modèle vise à atténuer ce problème de latence par l'identification de leaders dans le cadre d'un réseau comportemental. Dans ce réseau, les utilisateurs sont connectés quand ils ont des comportements de navigation semblables.

A la différence des études relatives à la détection de leaders citées dans la section 2.1, notre modèle utilise d'une part la structure topologique du réseau comportemental afin de déterminer des leaders potentiels. D'autre part, ce modèle repose sur la capacité à propager des avis ou des recommandations pertinentes pour l'identification de leaders fiables.

Les résultats présentés ici montrent l'intérêt de notre modèle pour la détection de leaders fiables dans le contexte des réseaux comportementaux. En effet, en plus de leur forte connectivité dans ces réseaux, ces leaders ont une importante potentialité de prédiction au vu de l'importante précision des appréciations propagées aux autres utilisateurs. Ils représentent ainsi le point d'entrée dans le réseau comportemental pour la recommandation de la nouveauté, ce qui permet d'atténuer le problème de latence.
De plus, notre modèle contribue à l'amélioration de la qualité des recommandations. En effet, comparé au FCS et en prenant en compte l'ensemble des items recommandés par les leaders, notre modèle génère des recommandations dont la précision est élevée.
Néanmoins, en considérant les prédictions générées par ces leaders dans cette expérimentation, notre modèle fait face au problème de couverture. En effet, seuls les TopN leaders comportementaux sont impliqués à la génération des recommandations. De ce fait, si le nombre de ces TopN leaders est restreint (comme nous l'avons choisi dans notre expérimentation), l'enjeu serait de trouver le compromis entre l'amélioration de la précision des prédictions et l'augmentation de la couverture.

De plus, au niveau de la fonction de propagation des appréciations, l'attribution des poids α est adaptée ici au corpus du Crédit Agricole et aux valeurs de similarités calculées dans le cadre de ce corpus. De ce fait, il serait judicieux d'opter pour une fonction de propagation où le poids α est automatiquement adaptatif selon les distributions des valeurs de similarités entre utilisateurs. En effet, nous pouvons avoir un cas par exemple, où tous les leaders comportementaux potentiels sont très connectés aux autres utilisateurs, mais dont les valeurs de similarité avec leurs voisins ne sont pas très élevés. De ce fait, l'application des poids α comme nous l'avons précisé dans cette expérimentation, n'est pas appropriée. Ainsi, le poids α devrait être dynamiquement ajusté et adapté aux valeurs de similarités calculées selon le corpus utilisé.

En outre, dans cette expérimentation, nous avons choisi de sélectionner les TopN10 et les TopN20 correspondant aux leaders potentiels, afin d'effectuer la propagation d'appréciations. Ce choix relève du fait que notre modèle vise à retrouver un petit segment d'utilisateurs représentatifs parmi l'ensemble d'utilisateurs, capables de prédire efficacement les appréciations des voisins. De plus, il s'agit de sélectionner les leaders potentiels dont la précision peut être évaluée (i.e. les leaders disposant d'items positivement appréciés dans le corpus test).

Par ailleurs, il serait intéressant d'étudier la qualité des nouveaux items recommandés par les leaders. En effet, l'expérimentation présentée ici nous a permis d'évaluer la qualité des appréciations propagées afin de déterminer les leaders fiables, ce qui signifie que nous disposions déjà (dans le corpus) des avis de ces leaders concernant ces items.
Une expérimentation complémentaire consistera ainsi à évaluer si tous les nouveaux items introduits au système de recommandation du portail Extranet du Crédit Agricole, ont été correctement recommandés par les leaders aux autres utilisateurs. En d'autres termes, il s'agit d'évaluer le retour des utilisateurs de l'Extranet par rapport à la pertinence des nouveaux items qui leur sont recommandés.

Conclusion et Perspectives

L'expansion de l'Internet et du nombre d'applications basées sur le Web tels que les portails d'entreprise, est associée à une prolifération d'information ou d'items dont le volume ne cesse de croître. Devant cette profusion et cette surcharge d'items, l'utilisateur peine à repérer l'information pertinente qui correspond le plus à ses besoins. Dans ce contexte, les systèmes de recommandation ont été développés en vue de faciliter l'accès à ces items pertinents. Leur objectif est d'anticiper les besoins de l'utilisateur en lui fournissant des recommandations d'items jugés pertinents par rapport à ses goûts.

Il existe une variété de techniques de recommandation parmi lesquelles le Filtrage Collaboratif (FC), qui constitue la technique la plus populaire. Le principe du FC consiste à retrouver des utilisateurs ayant des goûts similaires à ceux d'un utilisateur actif (ses voisins) et à utiliser leurs avis dans le but de lui recommander des items susceptibles de l'intéresser.

La dernière décennie a été marquée par un large déploiement des systèmes de recommandation exploitant notamment le FC, dans différents champs d'application intégrant les sites de e-commerce (e.g. Amazon), les sites de recrutement (e.g. JobFinder), les sites de musique (e.g. LastFM), etc.
Malgré cet engouement pour les systèmes de recommandation, certaines questions restent encore soulevées. L'une de ces questions est liée au manque de données, notamment le manque de notes explicites attribuées par des utilisateurs aux items. En effet, un système fondé sur le FC exploite ces notes afin d'évaluer les similarités entre utilisateurs en exploitant les items co-notés. Ces similarités permettent d'identifier les voisins dont les appréciations sont combinées pour calculer les recommandations. Or, si ces notes s'avèrent insuffisantes, le système sera incapable d'identifier un nombre significatif de voisins fiables.

Un autre enjeu pour les systèmes de recommandation est de résoudre le problème de démarrage à froid concernant la nouveauté d'un utilisateur et/ou d'un item. En l'absence des notes de la part de cet utilisateur et/ou sur cet item, il devient impossible pour le processus de filtrage de les intégrer dans les recommandations.

En outre, la précision des recommandations est un défi majeur pour tout système de recommandation dans la mesure où la pertinence des items recommandés permet de

contribuer à la satisfaction des attentes de l'utilisateur et à sa fidélisation au service en question.

A partir de ces questions de recherche et en prenant en compte le contexte d'un portail Extranet d'entreprise, nous avons proposé dans cette thèse de nouvelles approches de recommandation s'appuyant sur l'observation du comportement et sur l'analyse des usages des utilisateurs. L'objectif est d'améliorer l'usage des items accessibles sur ce portail, auprès des utilisateurs du Groupe Crédit Agricole.

Nous avons proposé un nouveau modèle comportemental de recommandation nommé BNCF, inspiré du Web Usage Mining et du FC. Ce modèle vise à modéliser les utilisateurs en analysant le comportement de navigation à partir des traces d'usage. Nous considérons en effet que deux utilisateurs ayant des motifs d'usage communs sont similaires.
Les similarités de comportement sont évaluées sur la base d'une mesure que nous avons proposée, qui tient notamment compte de la longueur maximale de motifs d'usage communs entre utilisateurs. Ces similarités sont par la suite exploitées afin d'identifier les voisins et générer des prédictions.
L'évaluation de la performance du système de recommandation montre que le BNCF contribue à une amélioration de la précision au niveau des items réellement recommandés par le système. Nous pouvons déduire que les traces d'usage sont une source d'information fiable permettant au système de recommandation de modéliser efficacement les utilisateurs (sans faire appel aux données de notes) et de générer des prédictions pertinentes.

Dans l'objectif d'améliorer davantage la performance du BNCF et de réduire l'espace de recherche des voisins, nous avons proposé une extension du BNCF à travers le modèle BNCF-PCS qui intègre une phase de clustering d'utilisateurs. Ce clustering a pour particularité de générer des clusters en considérant les similarités de voisins. L'avantage d'une telle démarche de clustering est la considération d'items supplémentaires (tous les items consultés par les utilisateurs) et non pas uniquement des items co-notés par les utilisateurs.
Les similarités de comportement navigationnel sont par la suite calculées dans chaque cluster généré en prenant en compte uniquement les séquences positives de navigation des utilisateurs (i.e. les séquences d'items positivement appréciés).
L'évaluation de ce modèle a permis de souligner une amélioration importante de la précision des recommandations, ainsi qu'une réduction du temps de calcul des similarités grâce à l'exploitation des clusters et à l'utilisation des séquences positives.
Néanmoins, malgré la contribution de cette démarche de clustering à la performance du système de recommandation, elle risque de négliger certaines informations pertinentes pendant le processus de réduction de l'espace de recherche. En effet, si un utilisateur n'a pas beaucoup de voisins communs avec les autres utilisateurs, le système trouvera des difficultés à lui retrouver des voisins fiables et à lui générer des recommandations pertinentes.
Ce constat nous a mené à une autre réflexion visant à remédier à ce problème de perte d'information ainsi qu'au problème de manque de données.
Il s'agit d'améliorer le processus d'identification des voisins, notamment par la recherche

de nouveaux liens entre utilisateurs. C'est dans cette optique que nous nous sommes inspirés des approches issues de l'analyse des réseaux sociaux pour prédire les liens pouvant relier les utilisateurs.
Ainsi, dans le cadre du modèle proposé D-BNCF, nous avons exploité l'information comportementale afin de modéliser les liens entre utilisateurs à travers un réseau comportemental. Nous avons proposé d'appliquer par la suite des méthodes de prédiction de lien et des associations transitives afin de densifier le réseau construit et découvrir de nouveaux voisins pour chaque utilisateur. Ces voisins sont impliqués dans le calcul des recommandations dans le but d'améliorer la qualité des recommandations ainsi que la capacité prédictive du système.
L'expérimentation met en évidence l'intérêt d'utiliser les nouveaux liens découverts par certaines méthodes de prédiction de lien. En effet, ces méthodes ont contribué à une meilleure précision des recommandations.

En outre, nous nous sommes intéressés à la question de démarrage à froid liée en particulier à la nouveauté d'un item (i.e. problème de latence). Nous avons ainsi proposé un modèle qui repose sur l'identification de leaders comportementaux pour la recommandation de la nouveauté. Nous considérons qu'un leader comportemental est un utilisateur connecté à un grand nombre d'utilisateurs ayant un comportement similaire et qui prédit fiablement les appréciations de ces utilisateurs.
Dans le but de détecter les leaders, notre modèle mesure d'abord la connectivité des utilisateurs pour déterminer des leaders potentiels. Par la suite, ce modèle évalue leur capacité à propager des recommandations pertinentes dans le but de déterminer les leaders les plus fiables.
Ainsi, en connaissant au préalable leurs opinions sur les nouveaux items, ces leaders constituent les utilisateurs représentatifs du réseau que le système doit cibler pour prédire les avis des autres utilisateurs sur ces nouveaux items.
L'évaluation de ce modèle a montré l'avantage de la propagation des avis des leaders pour la recommandation de la nouveauté. En effet, en prenant en compte l'ensemble des items recommandés par les leaders, notre modèle parvient à améliorer la qualité des recommandations.

Par ailleurs, en collaboration avec la société Sailendra S.A.S[23], les algorithmes développés autour du filtrage collaboratif comportemental (BNCF) ont été intégrés au niveau de la plate forme CASA du portail Extranet du Groupe Crédit Agricole (cf. section 1.3.2, chapitre 1, partie 2). Actuellement, ces algorithmes sont déployés et testés au niveau du site Extranet du Pôle Innovation avant d'être fonctionnels au niveau de tout le Groupe Crédit Agricole. Il est question d'intégrer également par la suite les autres modèles proposés dans le cadre de cette thèse.

[23]http ://www.sailendra.fr/

Perspectives

Notre travail de recherche ouvre des perspectives à court terme et à moyen et long terme.

A court terme

Nous souhaitons avoir **un retour d'expérience de la part des utilisateurs du Groupe Crédit Agricole S.A** concernant les recommandations qui leur sont proposées. Ces retours vont nous permettre d'évaluer directement l'intérêt de nos modèles pour la recommandation d'items pertinents et d'évaluer la satisfaction des utilisateurs. **Ces retours peuvent même être exploités par le système de recommandation en vue d'affiner les profils utilisateurs**.

L'un des objectifs que nous nous sommes fixés aussi pour les travaux futurs à court terme est d'élaborer **un modèle de recommandation qui ne requiert pas de notes pendant tout le processus de recommandation**. En effet, dans les modèles que nous avons proposés dans cette thèse, mêmes si les notes n'étaient pas exploitées en phase d'apprentissage, elles étaient souvent nécessaires dans la phase de prédiction.
Nous pouvons ainsi soit prendre en compte d'autres critères permettant de déterminer l'appréciation d'un item dans la phase de prédiction ou bien de considérer uniquement l'action de consulter ou pas un item dans cette même phase.

En outre, nous prévoyons d'étudier également **l'intérêt des liens sociaux pour les systèmes de recommandation** (i.e. les liens issus des relations sociales telle que la collaboration professionnelle ou l'amitié dans le cadre des plates-formes du Web social) et d'examiner jusqu'à quel point **ils peuvent être complémentaires avec les liens comportementaux**. Il s'agit d'évaluer l'impact de cette combinaison sur le choix des voisins et sur la performance du système de recommandation d'une manière générale.

A moyen et long terme

Dans le cadre de nos perspectives de recherche à moyen et long terme, nous envisageons d'étudier davantage **l'apport du leadership dans le cadre des systèmes de recommandation**. En effet, à notre connaissance peu d'études sont consacrées à l'identification de leaders dans ce cadre.
Nous souhaitons ainsi exploiter **les techniques issues de l'analyse de réseaux sociaux basées sur des approches topologiques ainsi que la technique d'analyse de contenu**. Il s'agit d'examiner notamment si l'hybridation des deux types de techniques permet la découverte de leaders pertinents. Les leaders peuvent ainsi être détectés en fonction de leur connectivité dans le réseau (construit par exemple sur la base de l'information comportementale) mais aussi sur la base de l'analyse du contenu des échanges (à travers des forums par exemple) qu'ils peuvent avoir avec les autres utilisateurs du réseau.
Par ailleurs, il semble prometteur d'étudier **l'applicabilité des techniques de sondage**

d'opinion dans le contexte des systèmes de recommandation. En effet l'objectif des techniques de sondage consiste à interroger un échantillon représentatif d'une population afin de déterminer l'opinion publique relative à tel ou tel sujet. Il est question de s'intéresser en particulier aux critères utilisés pour le choix de cet échantillon que nous pouvons considérer comme groupe de leaders dans notre contexte.

De plus, nous prévoyons d'étudier dans nos travaux futurs **la possibilité d'application du principe de combinaison de ressorts (principe des "séries parallèles"**) dans le cadre d'un réseau d'utilisateurs pour la recherche des liens qui les relient. Ce réseau peut être construit sur la base de l'information comportementale ou bien sur un autre type d'information. L'objectif de l'application du principe des "séries parallèles" est de rechercher tous les chemins pouvant relier deux nœuds donnés à travers ce réseau. Autrement dit, il s'agit d'examiner jusqu'à quel point ce principe peut être considéré comme une méthode de prédiction de lien dans un réseau d'utilisateurs, permettant la découverte de nouveaux voisins et remédiant au manque de données.

Poursuivre l'étude du problème de passage à l'échelle fait partie également de nos perspectives de recherche à moyen et long terme. Nous souhaitons ainsi limiter le nombre d'items et/ou d'utilisateurs afin de réduire l'espace de recherche de voisins dans le cadre du système de recommandation. Selon le contexte du portail Extranet du Crédit Agricole, le nombre d'utilisateurs (internes) reste relativement stable. Or, si nous considérons une dimension plus importante liée à un autre contexte (par exemple les sites de e-commerce caractérisés par un nombre élevé d'utilisateurs), **l'utilisation des techniques de réduction de dimensionnalité peut être envisagée à ce niveau, notamment les techniques de SVD**.

En outre, nous souhaitons aussi aborder **le problème du contexte utilisateur**. Le contexte est lié à l'environnement d'interaction de l'utilisateur avec le système (contexte professionnel ou personnel par exemple). L'enjeu est de développer des services de personnalisation proposant à l'utilisateur à tout moment et sur le bon support, des recommandations adaptées à son contexte spécifique, ce qui est susceptible d'améliorer sa satisfaction et sa fidélisation.

L'étude de l'évolution des goûts dans le temps s'inscrit également dans le cadre de nos perspectives de recherche. En effet, les appréciations des utilisateurs ont tendance à évoluer dans le temps. Ainsi, notre objectif est de proposer un système de recommandation capable de détecter le changement du comportement de l'utilisateur et d'adapter dynamiquement les recommandations en fonction des nouveaux besoins de cet utilisateur.

Table des figures

Liste des tableaux

Bibliographie

[Abhinandan *et al.*, 2007] Abhinandan, S. D. ; Mayur, D. ; Ashutosh, G. et Shyam, R. (2007). Google news personalization : scalable online collaborative filtering. *In Proceedings of the 16th international conference on World Wide Web*. ACM.

[Adamic et Adar, 2003] Adamic, L. et Adar, E. (2003). Friends and neighbors on the web. *Social networks*, 25(3):211–230.

[Adomavicius et Tuzhilin, 2005] Adomavicius, G. et Tuzhilin, A. (2005). Toward the next generation of recommender systems : A survey of the state-of-the-art. *IEEE transactions on knowledge and data engineering*, 17(6):734–749.

[Agarwal *et al.*, 2008] Agarwal, N. ; Liu, H. ; Tang, L. et Yu, P. (2008). Identifying the influential bloggers in a community. *In Proceedings of the international conference on Web search and web data mining (WSDM'08)*, pages 207–218, New York, NY, USA. ACM.

[Aggarwal *et al.*, 1999] Aggarwal, C. ; Wolf, J. ; Wu, K. et Yu, P. (1999). Horting hatches an egg : A new graph-theoretic approach to collaborative filtering. *In Proceedings of the ACM KDD Conference*. ACM.

[Agrawal *et al.*, 1993] Agrawal, R. ; Imieliński, T. et Swami, A. (1993). Mining association rules between sets of items in large databases. *In Proceedings of the 1993 ACM SIGMOD international conference on Management of data (SIGMOD'93)*, pages 207–216, New York, NY, USA. ACM.

[Agrawal et Srikant, 1994] Agrawal, R. et Srikant, R. (1994). Fast algorithms for mining association rules. *In Proceedings of VLDB Conference*, pages 487–499.

[Agrawal et Srikant, 1995] Agrawal, R. et Srikant, R. (1995). Mining sequential patterns. *In Proceedings of the 11th International Conference on Data Engineering*, pages 3–14.

[Aha *et al.*, 2000] Aha, D. ; Breslow, L. et Muoz-Avila, H. (2000). Conversational case-based reasoning. *Applied Intelligence*, (14):9–32.

[Anand et Mobasher, 2005] Anand, S. et Mobasher, B. (2005). Intelligent techniques for web personalization. *Lecture Notes in Artificial Intelligence*, 3169:1–36.

[Ayres *et al.*, 2002] Ayres, J. ; Flannick, J. ; Gehrke, J. et Yiu, T. (2002). Sequential pattern mining using a bitmap representation. *In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'02)*, pages 429–435, New York, NY, USA. ACM.

[Balabanović et Shoham, 1997] Balabanović, M. et Shoham, Y. (1997). Fab : content-based, collaborative recommendation. *Commun. ACM*, 40(3):66–72.

[Baltrunas et Ricci, 2007] Baltrunas, L. et Ricci, F. (2007). Dynamic item weighting and selection for collaborative filtering. *In Web mining 2.0 Workshop, ECML-PKDD 2007*. Springer-Verlag.

[Banerjee et Ghosh, 2001] Banerjee, A. et Ghosh, J. (2001). Clickstream clustering using weighted longest common subsequences. *In Proceedings of the Web Mining Workshop at the 1st SIAM Conference on Data Mining*.

[Barabási et Albert, 1999] Barabási, A. et Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509.

[Barabási *et al.*, 2002] Barabási, A. L. ; Jeong, H. ; Neda, Z. ; Ravasz, E. ; Schubert, A. et Vicsek, T. (2002). Evolution of the social network of scientific collaboration. *Physica A*, 311(3-4):590–614.

[Bartal *et al.*, 2009] Bartal, A. ; Sasson, E. et Ravid, G. (2009). Predicting links in social networks using text mining and sna. *In Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE.

[Basilico et Hofmann, 2004] Basilico, J. et Hofmann, T. (2004). A joint framework for collaborative and content filtering. *In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'04)*, pages 550–551, New York, USA. ACM.

[Baumgarten *et al.*, 2000] Baumgarten, M. ; Buchner, A. ; Anand, S. ; Mulvenna, M. et Hughes, J. (2000). *User-driven navigation pattern discovery from internet data*, chapitre Web Usage Analysis and User Profiling, pages 74–91. Lecture Notes in Computer Science. Springer-Verlag.

[Bell *et al.*, 2007] Bell, R. ; Yehuda, K. et Volinsky, K. (2007). Improved neighborhood-based collaborative filtering. *In KDDCup'07*.

[Bertrand-Pierron, 2006] Bertrand-Pierron, Y. (2006). Transfert de technologies sur le filtrage collaboratif : intégration des techniques de filtrage collaboratif sur un portail de gestion de contenu. Mémoire de D.E.A., UHP University Nancy 1.

[Billsus *et al.*, 2002] Billsus, D. ; Brunk, C. ; Evans, C. ; Gladish, B. et Pazzani, M. (2002). Adaptive interfaces for ubiquitous web access. *Communications of ACM*, 45(5):34–38.

[Billsus et Pazzani, 2000] Billsus, D. et Pazzani, M. (2000). User modeling for adaptive news access. *User-Modeling and User-Adapted Interaction*, 10(2-3):147–180.

[Billsus *et al.*, 2000] Billsus, D. ; Pazzani, M. et Chen, J. (2000). A learning agent for wireless news access. *In Proceedings of the 5th international conference on Intelligent user interfaces (IUI'00)*, pages 33–36, New York, NY, USA. ACM.

[Bodendorf et Kaiser, 2009] Bodendorf, F. et Kaiser, C. (2009). Detecting opinion leaders and trends in online social networks. *In Proceedings of the 2nd ACM workshop on Social web search and mining (SWSM'09)*, pages 65–68, New York, USA. ACM.

[Bonnin *et al.*, 2009] Bonnin, G. ; Brun, A. et Boyer, A. (2009). A low-order markov model integrating long-distance histories for collaborative recommender systems. *In Proceedings of the 13th international conference on Intelligent user interfaces (IUI'09)*, pages 57–66, New York, NY, USA. ACM.

[Brazma *et al.*, 1998] Brazma, A. ; Jonassen, I. ; Eidhammer, I. et Gilbert, D. (1998). Approaches to the automatic discovery of patterns in biosequences. *Journal of Computational Biology*, 5(2):277–304.

[Breese *et al.*, 1998] Breese, J. ; Heckerman, D. et Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. *In Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, pages 43–52. Morgan Kaufmann.

[Brin et Page, 1998] Brin, S. et Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *In Computer networks and ISDN systems*, pages 107–117. Elsevier Science Publishers B.V.

[Burke, 2000] Burke, R. (2000). Knowledge-based recommender systems. *Encyclopedia of Library and Information Systems*, 69(32).

[Burke, 2002] Burke, R. (2002). Hybrid recommender systems : Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370.

[Burke *et al.*, 1997] Burke, R. ; Hammond, K. et Young, B. (1997). The findme approach to assisted browsing. *IEEE Expert : Intelligent Systems and Their Applications*, 12(4):32–40.

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

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

[Chen *et al.*, 2009] Chen, J. ; Zaiane, O. R. et Goebel, R. (2009). Local community identification in social networks. *In Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM).* IEEE.

[Chen *et al.*, 1996] Chen, M. ; Hun, J. et Yu, P. (1996). Data mining : An overview from database perspective. *IEEE Transactions on Knowledge and Data Engineering*, 8:866–883.

[Cheon et Lee, 2005] Cheon, H. et Lee, H. (2005). *Opinion Leader Based Filtering*, volume 3815/2005 de *Lecture Notes in Computer Science.* Springer Berlin/Heidelberg.

[Cheype, 2006] Cheype, A. (2006). Recherche de motifs séquentiels pour guider l'interprétation des traces d'apprentissage. *In Actes des 1ères Rencontres Jeunes Chercheurs en EIAH (RJC-EIAH'2006)*, pages 123–130.

[Claypool *et al.*, 1999] Claypool, M. ; Gokhale, A. ; Miranda, T. ; Murnikov, P. ; Netes, D. et Sartin, M. (1999). Combining content-based and collaborative filters in an online newspaper. *In Proceedings of ACM SIGIR '99 Workshop on Recommender Systems : Algorithms and Evaluation.*

[Claypool *et al.*, 2001] Claypool, M. ; Le, P. ; Waseda, M. et Brown, D. (2001). Implicit interest indicators. *In Proceedings of ACM Intelligent User Interfaces Conference*.

[Coleman *et al.*, 1966] Coleman, J. ; Menzel, H. et Katz, E. (1966). *Medical Innovations : A Diffusion Study*. Bobbs-Merrill Co.

[Conner et Herlocker, 1999] Conner, M. et Herlocker, J. (1999). Clustering items for collaborative filtering. *In Proceedings of the ACM SIGIR Workshop on Recommender Systems*.

[Cooke, 2006] Cooke, R. (2006). *Link prediction and link detection in sequences of large social networks using temporal and local metrics*. Thèse de doctorat, University of cape Town.

[Cooley *et al.*, 1999] Cooley, R. ; Mobasher, B. et Srivastava, J. (1999). Data preparation for mining world wide web browsing patterns. *Journal of Knowledge and Information Systems*, 1(1):5–32.

[Cosley *et al.*, 2002] Cosley, D. ; Lawrence, S. et Pennock, D. (2002). Referee : An open framework for practical testing of recommender systems using researchindex. *In Proceedings of the 28th international conference on Very Large Data Bases*, page 46. VLDB Endowment.

[Crandall *et al.*, 2008] Crandall, D. ; Cosley, D. ; Huttenlocher, D. ; Kleinberg, J. et Suri, S. (2008). Feedback effects between similarity and social influence in online communities. *In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM.

[Domingos et Richardson, 2001] Domingos, P. et Richardson, M. (2001). Mining the network value of customers. *In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'01)*, pages 57–66, New York, NY, USA. ACM.

[Doyle et Cunningham, 2000] Doyle, M. et Cunningham, P. (2000). A dynamic approach to reducing dialog in on-line decision guides. *In Proceedings of the 5th European Workshop on Advances in Case-Based Reasoning (EWCBR'00)*, pages 49–60, London, UK. Springer-Verlag.

[Eirinaki *et al.*, 2005] Eirinaki, M. ; Vazirgiannis, M. et Kapogiannis, D. (2005). Web path recommendations based on page ranking and markov models. *In Proceedings of the 7th annual ACM international workshop on Web information and data management*. ACM Press.

[Esslimani *et al.*, 2008a] Esslimani ; Brun, A. et Boyer, A. (2008a). Behavioral similarities for collaborative recommendations. *Journal of Digital Information Management*, 6(6):442–448.

[Esslimani *et al.*, 2008b] Esslimani, I. ; Brun, A. et Boyer, A. (2008b). Enhancing collaborative filtering by frequent usage patterns. *In Proceedings of the First IEEE International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2008). Workshop on Recommender Systems and Personalized Retrieval*, pages 180–185.

[Esslimani *et al.*, 2009a] Esslimani, I. ; Brun, A. et Boyer, A. (2009a). A collaborative filtering approach combining clustering and navigational based correlations. *In Proceedings of the 5th International Conference on Web Information Systems and Technologies (WEBIST 2009)*, pages 364–369. INSTICC.

[Esslimani *et al.*, 2009b] Esslimani, I. ; Brun, A. et Boyer, A. (2009b). From social networks to behavioral networks in recommender systems. *In Proceedings of The 2009 International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 143–148. IEEE Computer society.

[Esslimani *et al.*, 2009c] Esslimani, I. ; Brun, A. et Boyer, A. (2009c). Vers l'exploitation de la transitivité dans les réseaux comportementaux pour les systèmes de recommandations. *In 7ème colloque du chapitre français de l'ISKO sur l'Intelligence collective et l'organisation des connaissances*.

[Esslimani *et al.*, 2010a] Esslimani, I. ; Brun, A. et Boyer, A. (2010a). Densifying a behavioral recommender system by social networks link prediction methods. *The Social Network Analysis and Mining Journal*.

[Esslimani *et al.*, 2010b] Esslimani, I. ; Brun, A. et Boyer, A. (2010b). Detecting leaders in behavioral networks. *In Proceedings of The 2010 International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE Computer society.

[Esslimani *et al.*, 2010c] Esslimani, I. ; Brun, A. et Boyer, A. (2010c). Detecting leaders to alleviate latency in recommender systems. *In Proceedings of the EC-WEB 2010 Conference*, pages 229–240. Springer-Verlag.

[Freyne *et al.*, 2007] Freyne, J. ; Farzan, R. et Coyle, M. (2007). Toward the exploitation of social access patterns for recommendation. *In Proceedings of the 2007 ACM conference on Recommender systems*. ACM.

[Fu *et al.*, 2000] Fu, X. ; Budzik, J. et Hammond, K. (2000). Mining navigation history for recommendation. *In Proceedings of the 5th international conference on Intelligent User Interfaces (IUI'00)*, pages 106–112. ACM.

[Gaul et Schmidt-Thieme, 2001] Gaul, G. et Schmidt-Thieme, L. (2001). Frequent substructures in web usage data : A unified approach. *In Proceedings of Web Mining Workshop, First SIAM International Conference on Data Mining 2001 (ICDM)*.

[George et Merugu, 2005] George, T. et Merugu, S. (2005). A scalable collaborative filtering framework based on co-clustering. *In Proceedings of the Fifth IEEE International Conference on Data Mining*. IEEE Computer Society.

[Gery et Haddad, 2003] Gery, M. et Haddad, H. (2003). Evaluation of web usage mining approaches for user's next request prediction. *In Proceedings of the 5th ACM international workshop on Web information and data management*. ACM Press.

[Gladwell, 2000] Gladwell, M. (2000). *The Tipping Point : How Little Things Can Make a Big Difference*. Little Brown, New York.

[Golbeck, 2009] Golbeck, J. (2009). Trust and nuanced profile similarity in online social networks. *ACM Transactions on the WEB (TWEB)*, 3(4):1–33.

[Goldberg *et al.*, 1992] Goldberg, D. ; Nichols, D. ; Oki, B. et Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70.

[Goldberg *et al.*, 2001] Goldberg, K. ; Roeder, T. ; Gupta, D. et Perkins, C. (2001). Eigentaste : A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151.

[Gong *et al.*, 2009] Gong, S. ; Ye, H. et Dai, Y. (2009). Combining singular value decomposition and item-based recommender in collaborative filtering. *In Proceedings of the 2009 Second International Workshop on Knowledge Discovery and Data Mining (WKDD'09)*, pages 769–772, Washington, DC, USA. IEEE Computer Society.

[Good *et al.*, 1999] Good, N. ; Schafer, J. ; Konstan, J. ; Borchers, A. ; Sarwar, B. ; Herlocker, J. et Riedl, J. (1999). Combining collaborative filtering with personal agents for better recommendations. *In Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence (AAAI'99/IAAI'99)*, pages 439–446, Menlo Park, CA, USA. American Association for Artificial Intelligence.

[Goyal *et al.*, 2008] Goyal, A. ; Bonchi, F. et Lakshmanan, L. (2008). Discovering leaders from community actions. *In Proceeding of the 17th ACM conference on Information and knowledge management (CIKM'08)*, pages 499–508, New York, NY, USA. ACM.

[Grcar, 2004] Grcar, M. (2004). User profiling : Collaborative filtering. *In Proceedings of the conference on data mining and warehouses (SIKDD 2004) at multiconference IS 2004*.

[Han et Kamber, 2001] Han, J. et Kamber, M. (2001). *Data Mining : Concepts and Techniques*. Morgan Kaufmann, San Francisco, California, USA.

[Han *et al.*, 2000] Han, J. ; Pei, J. ; Mortazavi-Asl, B. ; Chen, Q. ; Dayal, U. et Hsu, M. (2000). Freespan : frequent pattern-projected sequential pattern mining. *In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'00)*, pages 355–359, New York, NY, USA. ACM.

[Hao *et al.*, 2007] Hao, M. ; King, I. et Lyu, M. R. (2007). Effective missing data prediction for collaborative filtering. *In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM.

[Herlocker *et al.*, 1999] Herlocker, J. ; Konstan, J. ; Borchers, A. et Riedl, J. (1999). An algorithmic framework for performing collaborative filtering. *In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*.

[Herlocker *et al.*, 2004] Herlocker, J. ; Konstan, J. ; Terveen, L. et Riedl, J. (2004). Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1): 5–53.

[Hofmann, 2003] Hofmann, T. (2003). Gaussian latent semantic models for collaborative filtering. *In Proceedings of the 26th Annual International ACM SIGIR Conference*.

[Hofmann, 2004] Hofmann, T. (2004). Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems (TOIS)*, 22(1):89–115.

[Hopfield, 1982] Hopfield, J. (1982). Neural network and physical system with emergent collective computational abilities. *Nat.Acad.Sci*, 79:2554–2558.

[Hu et Panda, 2004] Hu, Y. et Panda, B. (2004). A data mining approach for database intrusion detection. *In Proceedings of the 2004 ACM symposium on Applied computing (SAC'04)*, pages 711–716, New York, NY, USA. ACM.

[Huang *et al.*, 2004] Huang, Z. ; Chen, H. et Zeng, D. (2004). Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. *ACM Transactions on Information Systems (TOIS)*, 22(1):116–142.

[Huang *et al.*, 2002] Huang, Z. ; Chung, W. ; Ong, T. et Chen, H. (2002). A graph-based recommender system for digital library. *In Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*. ACM.

[Huang *et al.*, 2005] Huang, Z. ; Li, X. et Chen, H. (2005). Link prediction approach to collaborative filtering. *In Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*. ACM.

[Huang et Zeng, 2005] Huang, Z. et Zeng, D. (2005). Why does collaborative filtering work ? a recommendation model validation and selection by analyzing bipartite random graphs. *In Proceedings of Workshop of information Technologies and Systems*.

[Jalali *et al.*, 2008] Jalali, M. ; Mustapha, N. ; Sulaiman, N. et Mamat, A. (2008). A web usage mining approach based on lcs algorithm in online predicting recommendation systems. *In Proceedings of 12th conference of information visualisation*.

[Jamali et Abolhassani, 2006] Jamali, M. et Abolhassani, H. (2006). Different aspects of social network analysis. *In Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*.

[Jäschke *et al.*, 2007] Jäschke, R. ; Marinho, L. ; Hotho, A. ; Schmidt-Thieme, L. et Stumme, G. (2007). Tag recommendations in folksonomies. *Knowledge Discovery in Databases (PKDD 2007)*, pages 506–514.

[Jiang *et al.*, 2006] Jiang, X. ; Song, W. et Feng, W. (2006). Optimizing collaborative filtering by interpolating the individual and group behaviors. *In APWeb*.

[Katz et Lazarsfeld, 1955] Katz, E. et Lazarsfeld, P. (1955). *Personal Influence : the Part Played by People in the Flow of Mass Communications*. Free Press.

[Kaufman et Rousseuw, 1990] Kaufman, L. et Rousseuw, P. (1990). *Finding Groups in Data : An Introduction to Cluster Analysis*. John Wiley and Sons, New York.

[Kautz *et al.*, 1997] Kautz, H. ; Selman, B. et Shah, M. (1997). Referralweb : Combining social networks and collaborative filtering. *Communications of the ACM*, 30(3).

[Keller et Berry, 2003] Keller, E. et Berry, J. (2003). *The influentials*. Simon and Schuster Ed.

[Kempe *et al.*, 2003] Kempe, D. ; Kleinberg, J. et Tardos, E. (2003). Maximizing the spread of influence through a social network. *In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'03)*, pages 137–146, New York, NY, USA. ACM.

[Kim *et al.*, 2002] Kim, T.-H. ; Ryu, Y.-S. ; Park, S.-I. et Yang, S.-B. (2002). An improved recommendation algorithm in collaborative filtering. *E-Commerce and Web Technologies*, pages 517–529.

[Krulwich, 1997] Krulwich, B. (1997). Lifestyle finder : Intelligent user profiling using large-scale demographic data. *AI Magazine*, (18):37–45.

[Krulwich et Burkey, 1996] Krulwich, B. et Burkey, C. (1996). Learning user information interests through extraction of semantically significant phrases. *In Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access*. Stanford, CA.

[Lam et Riedl, 2004] Lam, S. et Riedl, J. (2004). Shilling recommender systems for fun and profit. *In Proceedings of the 13th international conference on World Wide Web (WWW'04)*, pages 393–402, New York, NY, USA. ACM.

[Lang, 1995] Lang, K. (1995). Newsweeder : Learning to filter netnews. *In Proceedings of the 12th International Conference on Machine Learning (ICML95)*, pages 331–339.

[Liben-Nowell et Kleinberg, 2003] Liben-Nowell, D. et Kleinberg, J. (2003). The link prediction problem for social networks. *In Proceedings of the 12th international conference on Information and knowledge management*. ACM.

[Lieberman, 1995] Lieberman, H. (1995). Letizia : An agent that assists web browsing. *In International Joint Conference on Artificial Intelligence*, pages 924–929.

[Lim *et al.*, 2003] Lim, M. ; Negnvitsky, M. et Hartnett, J. (2003). Artificial intelligence applications for analysis of e-mail communication activities. *In Proceedings of the International Conference On Artificial Intelligence In Science And Technology*.

[Lin *et al.*, 2002] Lin, W. ; Alvarez, S. et Ruiz, C. (2002). Efficient adaptive-support association rule mining for recommender systems. *Data Mining and Knowledge Discovery*, 6(1):83–105.

[Linden *et al.*, 2003] Linden, G. ; Smith, B. et York, J. (2003). Amazon.com recommendations : Item-to-item collaborative filtering. *IEEE Internet computing*, 7(1):76–80.

[Littlestone et Warmuth, 1994] Littlestone, N. et Warmuth, M. K. (1994). The weighted majority algorithm. *Inf. Comput.*, 108(2):212–261.

[Liu *et al.*, 2007] Liu, Y. ; Huang, X. et An, A. (2007). Personalized recommendation with adaptive mixture of markov models. *Journal of American Society for Information Science and Technology*, 58(12):1851–1870.

[MacQueen, 1967] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *In Proceedings of the 5th Symposium on Math, Statistics and Probability*, pages 281–297.

[Massa et Bhattacharjee, 2004] Massa, P. et Bhattacharjee, B. (2004). Using trust in recommender systems : an experimental analysis. *In Proceedings of 2nd International Conference on Trust Managment*.

[McGinty et Smyth, 2005] McGinty, L. et Smyth, B. (2005). *Intelligent techniques for web personalization*, volume 3169/2005 de *Lecture Notes in Computer Science*, chapitre Improving the performance of recommender systems that use critiquing, pages 114–132. Springer Berlin / Heidelberg.

[McLaughlin et Herlocker, 2004] McLaughlin, M. et Herlocker, J. (2004). A collaborative filtering algorithm and evaluation metric that accurately model the user experience. *In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'04)*, pages 329–336, New York, NY, USA. ACM.

[McNee *et al.*, 2002] McNee, S. ; Albert, I. ; Cosley, D. ; Gopalkrishnan, P. ; Lam, S. ; Rashid, A. ; Konstan, J. et Riedl, J. (2002). On the recommending of citations for research papers. *In Proceedings of the 2002 ACM conference on Computer supported cooperative work*, page 125. ACM.

[Mehta *et al.*, 2007] Mehta, B. ; Hofmann, T. et Nejdl, W. (2007). Robust collaborative filtering. *In Proceedings of the 2007 ACM conference on Recommender systems (RecSys'07)*, pages 49–56, New York, NY, USA. ACM.

[Melville *et al.*, 2002] Melville, P. ; Mooney, R. et Nagarajan, R. (2002). Content-boosted collaborative filtering for improved recommendations. *In Proceedings of the Eighteenth national conference on Artificial intelligence*, pages 187–192, Menlo Park, CA, USA. American Association for Artificial Intelligence.

[Middleton *et al.*, 2004] Middleton, S. ; Shadbolt, N. et Roure, D. D. (2004). Ontological user profiling in recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):54–88.

[Mislove *et al.*, 2007] Mislove, A. ; Marcon, M. ; Gummadi, K. P. ; Druschel, P. et Bhattacharjee, B. (2007). Measurement and analysis of online social networks. *In Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. ACM.

[Mladenic, 1999] Mladenic, D. (1999). Text-learning and related intelligent agents : A survey. *IEEE Intelligent Systems*, 14(4):44–54.

[Mobasher *et al.*, 2001] Mobasher, B. ; Dai, H. ; Luo, T. et Nakagawa, M. (2001). Improving the effectiveness of collaborative filtering on anonymous web usage data. *In Proceedings of the IJCAI 2001 Workshop on Intelligent Techniques for Web Personalization (ITWP01)*.

[Nakagawa et Mobasher, 2003] Nakagawa, M. et Mobasher, B. (2003). A hybrid web personalization model based on site connectivity. *In Proceedings of WebKDD Workshop at KDD'2003*, pages 59–70.

[Newman, 2001] Newman, M. (2001). Clustering and preferential attachment in growing networks. *Physical Review Letters*, 64(025102).

[Newman, 2003] Newman, M. (2003). The structure and function of complex networks. *SIAM Review*, 45:167–256.

[Nguyen *et al.*, 2006] Nguyen, A. ; Denos, N. et Berrut, C. (2006). Exploitation des données disponibles à froid pour améliorer le démarrage à froid dans les systèmes de filtrage d'information. *In Actes du XXIV Congrès d'INFORSID*, pages 81–95.

[Nichols, 1997] Nichols, D. (1997). Implicit rating and filtering. *In Proceedings of the Fifth DELOS Workshop on Filtering and Collaborative Filtering*, pages 31–36. ERCIM.

[O'Donovan et Smyth, 2005] O'Donovan, J. et Smyth, B. (2005). Trust in recommender systems. *In Proceedings of the 10th international conference on Intelligent user interfaces (IUI'05)*, pages 167–174, New York, NY, USA. ACM.

[Ohn *et al.*, 2003] Ohn, J. H. ; Kim, J. et Kim, J. H. (2003). Social network analysis of gene expression data. *In Proceedings of AMIA symposium : Biomedical and health informatics*. AMIA.

[O'Mahony *et al.*, 2006] O'Mahony, M. ; Hurley, N. et Silvestre, G. (2006). Detecting noise in recommender system databases. *In Proceedings of the 11th international conference on Intelligent user interfaces (IUI'06)*, pages 109–115, New York, NY, USA. ACM.

[O'Reilly, 2005] O'Reilly, T. (2005). What is web 2.0. design patterns and business models for the next generation of software. *In Proceedings of Web 2.0 Conference*.

[Papagelis *et al.*, 2005] Papagelis, M. ; Plexousakis, D. et Kutsuras, T. (2005). Alleviating the sparsity problem of collaborative filtering using trust inferences. *In iTrust*. Springer-Verlag Berlin Heidelberg.

[Paris *et al.*, 2009] Paris, C. ; Colineau, N. ; Thomas, P. et Wilkinson, R. (2009). Stakeholders and their respective costs-benefits in ir evaluation. *In SIGIR 2009 Workshop on the Future of IR Evaluation*.

[Park *et al.*, 2006] Park, S. ; Pennock, D. ; Madani, O. ; Good, N. et DeCoste, D. (2006). Naïve filterbots for robust cold-start recommendations. *In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'06)*, pages 699–705, New York, NY, USA. ACM.

[Pass *et al.*, 2006] Pass, G. ; Chowdhury, A. et Torgeson, C. (2006). A picture of search. *In Proceedings of the 1st international conference on Scalable information systems*.

[Pazzani et Billsus, 2007] Pazzani, M. et Billsus, D. (2007). *The Adaptive Web*, volume 4321/2007 de *Lecture Notes in Computer Science*, chapitre Content-Based Recommendation Systems, pages 325–341. Springer Berlin / Heidelberg.

[Pazzani, 1999] Pazzani, M. J. (1999). A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Revue*, 13(5-6):393–408.

[Pessiot *et al.*, 2006] Pessiot, J. ; Vinh, T. ; Usunier, N. ; Amini, M. et Gallinari, P. (2006). Factorisation en matrices non-négatives pour le filtrage collaboratif. *In Actes de CORIA 2006*.

[Popescul *et al.*, 2001] Popescul, A. ; Ungar, L. ; Pennock, D. M. et Lawrence, S. (2001). Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. *In Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence (UAI'01)*, pages 437–444, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Rafter *et al.*, 2000] Rafter, R. ; Bradley, K. et Smyth, B. (2000). *Adaptive Hypermedia and Adaptive Web-Based Systems*, volume 892/2000 de *Computer Science*, chapitre Automated Collaborative Filtering Applications for Online Recruitment Services, pages 363–368. Springer Berlin Heidelberg.

[Rashid *et al.*, 2008] Rashid, A. ; Karypis, G. et Riedl, J. (2008). Learning preferences of new users in recommender systems : an information theoretic approach. *SIGKDD Explor. Newsl.*, 10(2):90–100.

[Resnick *et al.*, 1994] Resnick, P. ; Iacovou, N. ; Suchak, M. ; Bergstrom, P. et Riedl, J. (1994). Grouplens : An open architecture for collaborative filtering of netnews. *In Proceedings of the ACM conference on computer-supported cooperative work*.

[Resnick et Varian, 1997] Resnick, P. et Varian, H. (1997). Recommender systems. *Communications of ACM*, 40(3):56–58.

[Salton, 1989] Salton, G. (1989). *Automatic text processing : the transformation, analysis, and retrieval of information by computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[Salton et McGill, 1983] Salton, G. et McGill, M. (1983). *Introduction to modern information retrieval*. McGraw-Hill, New York.

[Sarwar *et al.*, 2001] Sarwar, B. ; Karypis, G. ; Konstan, J. et Reidl, J. (2001). Item-based collaborative filtering recommendation algorithms. *In Proceedings of the 10th international conference on World Wide Web (WWW'01)*, pages 285–295, New York, NY, USA. ACM.

[Sarwar *et al.*, 2000a] Sarwar, B. ; Karypis, G. ; Konstan, J. et Riedl, J. (2000a). Analysis of recommendation algorithms for e-commerce. *In Proceedings of the 2nd ACM conference on Electronic commerce (EC'00)*, pages 158–167, New York, NY, USA. ACM.

[Sarwar *et al.*, 2000b] Sarwar, B. ; Karypis, G. ; Konstan, J. et Riedl, J. (2000b). Application of dimensionality reduction in recommender system - a case study. *In ACM WebKDD 2000 Web Mining for ECommerce Workshop*.

[Sarwar *et al.*, 2002] Sarwar, B. ; Karypis, G. ; Konstan, J. et Riedl, J. (2002). Recommender systems for large-scale e-commerce : Scalable neighborhood formation using clustering. *In Proceedings of the Fifth International Conference on Computer and Information Technology*, pages 158–167.

[Sarwar *et al.*, 1998] Sarwar, B. ; Konstan, J. ; Borchers, A. ; Herlocker, J. ; Miller, B. et Riedl, J. (1998). Using filtering agents to improve prediction quality in the grouplens research collaborative filtering system. *In Proceedings of the 1998 ACM conference on Computer supported cooperative work (CSCW'98)*, pages 345–354, New York, NY, USA. ACM.

[Schafer *et al.*, 2007] Schafer, J. ; Frankowski, D. ; Herlocker, J. et Sen, S. (2007). Collaborative filtering recommender systems. pages 291–324.

[Schein *et al.*, 2002] Schein, A. ; Popescul, A. ; Ungar, L. H. et Pennock, D. M. (2002). Methods and metrics for cold-start recommendations. *In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'02)*, pages 253–260, New York, USA. ACM.

[Shani *et al.*, 2005] Shani, G. ; Heckerman, D. et Brafman, R. (2005). An mdp-based recommender system. *The Journal of Machine Learning Research*, 6:1265–1295.

[Shardanand et Maes, 1995] Shardanand, U. et Maes, P. (1995). Social information filtering : algorithms for automating "word of mouth". *In Proceedings of the SIGCHI conference on Human factors in computing systems (CHI'95)*, pages 210–217, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co.

[Shimazu, 2001] Shimazu, H. (2001). Expertclerk : navigating shoppers' buying process with the combination of asking and proposing. *In Proceedings of the 17th international joint conference on Artificial intelligence (IJCAI'01)*, pages 1443–1448, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Smyth, 2007] Smyth, B. (2007). Case-based recommendation. *The adaptive web : methods and strategies of web personalization*, pages 342–376.

[Smyth et Cotter, 2000] Smyth, B. et Cotter, P. (2000). A personalized tv listings service for the digital tv age. *Knowledge-Based Systems*, (13):53–59.

[Soboroff et Nicholas, 1999] Soboroff, I. et Nicholas, C. (1999). Combining content and collaboration in text filtering. *In Proceedings of the IJCAI-99, Workshop on Machine Learning for Information Filtering*.

[Sollenborn et Funk, 2002] Sollenborn, M. et Funk, P. (2002). Category-based filtering and user stereotype cases to reduce the latency problem in recommender systems. *In Proceedings of the 6th European Conference on Advances in Case-Based Reasoning (ECCBR'02)*, pages 395–420, London, UK. Springer-Verlag.

[Srikant et Agrawal, 1996] Srikant, R. et Agrawal, R. (1996). Mining sequential patterns : Generalizations and performance improvements. *In Proceedings of the 5th International Conference on Extending Database Technology (EDBT'96)*, pages 3–17, London, UK. Springer-Verlag.

[Srivastava *et al.*, 2000] Srivastava, J. ; Cooley, R. ; Deshpande, M. et Tan, P.-N. (2000). Web usage mining : discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2):12–23.

[Su et Khoshgoftaar, 2009] Su, X. et Khoshgoftaar, T. (2009). A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, Janvier 2009:1–20.

[Svensson *et al.*, 2005] Svensson, M. ; Höök, K. et Cöster, R. (2005). Designing and evaluating kalas : A social navigation system for food recipes. *ACM Transactions on Computer-Human Interactions (TOCHI)*, 12(3):374–400.

[Tamine-Lechani et Calabretto, 2008] Tamine-Lechani, L. et Calabretto, S. (2008). *Recherche d'information : état des lieux et perspectives*, chapitre Recherche d'information contextuelle et Web, pages 201–224.

[Tang et McCalla, 2003] Tang, T. et McCalla, G. (2003). Mining implicit ratings for focused collaborative filtering for paper recommendations. *In 9th International Conference on User Modeling (UM 2003), Workshop on User and Group Models for Web-based Adaptive Collaborative Environments*.

[Tran, 2006] Tran, T. (2006). Designing recommender systems for e-commerce : an integration approach. *In Proceedings of the 8th international conference on Electronic commerce (ICEC'06)*, pages 512–518, New York, NY, USA. ACM.

[Tufféry, 2007] Tufféry, S. (2007). *Data mining et statistique décisionnelle : l'intelligence des données*. Editions Ophrys.

[Ungar et Foster, 1998] Ungar, L. et Foster, D. (1998). Clustering methods for collaborative filtering. *In Proceedings of the AAAI Workshop on Recommendation Systems*, pages 112–125.

[Valente, 1995] Valente, T. (1995). *Network models of the diffusion of innovations*. Hampton Press.

[Verma *et al.*, 2009] Verma, S. ; Patel, S. et Abhari, A. (2009). Adaptive web navigation. *In Proceedings of the 2009 Spring Simulation Multiconference (SpringSim'09)*, pages 1–4, San Diego, CA, USA. Society for Computer Simulation International.

[Viappiani *et al.*, 2006] Viappiani, P. ; Faltings, B. et Pu, P. (2006). Preference-based search using example-critiquing with suggestions. *Journal of artificial intelligence Research*, 27:465–503.

[Vozalis et Margaritis, 2006] Vozalis, M. et Margaritis, K. (2006). On the enhancement of collaborative filtering by demographic data. *Web Intelligence and Agent Systems : An International Journal (WIAS)*, 4(2):117–138.

[Wagner et Fischer, 1974] Wagner, R. et Fischer, M. (1974). The string-to-string correction problem. *Journal of the ACM (JACM)*, 21:168–173.

[Wang et Shao, 2004] Wang, F.-H. et Shao, H.-M. (2004). Effective personalized recommendation based on time-framed navigation clustering and association mining. 27(3):365–377.

[Wang *et al.*, 2008] Wang, Y. ; Dai, W. et Yuan, Y. (2008). Website browsing aid : A navigation graph-based recommendation system. *Decision Support Systems*, 45(3):387–400.

[Watts et Dodds, 2007] Watts, D. et Dodds, P. (2007). Influentials, networks, and public opinion formation. *Journal of Consumer Research*, 34(4):441–458.

[Webster et Vassileva, 2007] Webster, A. et Vassileva, J. (2007). Push-poll recommender system : Supporting word of mouth. *User Modeling 2007*, pages 278–287.

[Xiaoyuan *et al.*, 2007] Xiaoyuan, S. ; Russell, G. ; Taghi, M. et Xingquan, Z. (2007). Hybrid collaborative filtering algorithms using a mixture of experts. *In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE.

[Xue *et al.*, 2005] Xue, G. ; Lin, C. et Yang, Q. (2005). Scalable collaborative filtering using cluster-based smoothing. *In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*.

[Yamanishi *et al.*, 2005] Yamanishi, Y. ; Vert, J.-P. et Kanehisa, M. (2005). Supervised enzyme network inference from the integration of genomic data and chemical information. *Bioinformatics*, 21(1):468–477.

[Zaki, 2001] Zaki, M. (2001). Spade : An efficient algorithm for mining frequent sequences. *Machine Learning*, 42(1-2):31–60.

[Zhang *et al.*, 2005] Zhang, S. ; Wang, W. ; Ford, J. ; Makedon, F. et Pearlman, J. (2005). Using singular value decomposition approximation for collaborative filtering. *In Proceedings of the Seventh IEEE International Conference on E-Commerce Technology (CEC'05)*, pages 257–264, Washington, DC, USA. IEEE Computer Society.

[Zheng *et al.*, 2007] Zheng, R. ; Provost, F. et Ghose, A. (2007). Social network collaborative filtering. *IOMS : Information Systems Working Papers*, CeDER-07-04.

[Ziegler *et al.*, 2005] Ziegler, C. ; McNee, S. ; Konstan, J. et Lausen, G. (2005). Improving recommendation lists through topic diversification. *In Proceedings of the 14th international conference on World Wide Web (WWW'05)*, pages 22–32, New York, NY, USA. ACM.

[Zimdars *et al.*, 2001] Zimdars, A. ; Chickering, D. et Meek, C. (2001). Using temporal data for making recommendations. *In Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence (UAI'01)*, pages 580–588, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Résumé

Internet met à la disposition des utilisateurs une large variété d'items dont le volume est sans cesse croissant. Devant cette surcharge d'items, l'utilisateur peine à repérer les items qui correspondent à ses besoins. C'est dans ce contexte que les systèmes de recommandation se sont développés, dans la mesure où ils permettent de faciliter l'accès aux items susceptibles d'intéresser l'utilisateur. Néanmoins, malgré le succès des systèmes de recommandation, certaines questions de recherche restent soulevées telles que : le manque de données, l'identification de voisins fiables, la précision des recommandations et la recommandation de la nouveauté. En vue de répondre à ces questions, nous avons proposé à travers cette thèse une nouvelle approche de recommandation inspirée du web usage mining et du filtrage collaboratif. Cette approche repose sur l'observation du comportement de l'utilisateur et sur l'analyse de ses usages en vue de générer des recommandations. En outre, nous nous sommes inspirés des techniques utilisées dans le domaine de l'analyse des réseaux sociaux afin de prédire les liens à travers un réseau d'utilisateurs construit sur la base des similarités de comportement. L'objectif est de pallier le manque de données et d'améliorer l'identification de voisins fiables. De plus, dans la perspective d'atténuer le problème de démarrage à froid (concernant les nouveaux items), nous avons proposé une approche de recommandation qui repose sur la détection de leaders pour la recommandation de la nouveauté.

Mots-clés : systèmes de recommandation, filtrage collaboratif, analyse des usages, prédiction de lien, réseau comportemental, leadership

Abstract

The development of internet engendred an important proliferation of items. Thus, users are often overwhelmed and unable to detect the items corresponding to their needs. Therefore, the need of tools for automatic personalization of information becomes heightened. Recommender systems are widely used for this purpose thanks to their ability to guide users towards relevant items. Despite the success of recommender systems in many application areas, some research questions still remain. Some of these questions concern sparsity, selection of reliable neighbors, precision of recommendations and cold start problem. In this PhD thesis we explored these issues and proposed some solutions. We suggested a new approach inspired from web usage mining and collaborative filtering. This approach observes users' behavior and exploits usage analysis to generate recommendations. In addition, we applied link prediction methods, from social network analysis area, in order to predict new links in a behavioral network. The objective is to overcome sparsity and to improve neighbor selection. Moreover, with the perspective of alleviating the cold start problem (for new items), we proposed a recommendation approach based on leader detection. These leaders can propagate their appreciations towards their neighbors and predict accurately their future preferences.

Keywords : recommender systems, collaborative filtering, usage analysis, link prediction, behavioral network, leadership