Université de Montréal

# Papyres : Un système de gestion et de recommandation d'articles de recherche

par
Naak Amine

Département d'Informatique et de Recherche Opérationnelle
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de Maîtrise ès Sciences
en Informatique

Juillet, 2009

Université de Montréal
Faculté des arts et des sciences

Ce mémoire est intitulé :

# Papyres : Un système de gestion et de recommandation d'articles de recherche

présenté par :
Naak Amine

a été évalué par un jury composé des personnes suivantes :

Jian-Yun Nie, président-rapporteur
Esma Aïmeur, directrice de recherche
Philippe Langlais, membre du jury

# Résumé

Les étudiants gradués et les professeurs (les chercheurs, en général), accèdent, passent en revue et utilisent régulièrement un grand nombre d'articles, cependant aucun des outils et solutions existants ne fournit la vaste gamme de fonctionnalités exigées pour gérer correctement ces ressources. En effet, les *systèmes de gestion de bibliographie* gèrent les références et les citations, mais ne parviennent pas à aider les chercheurs à manipuler et à localiser des ressources. D'autre part, les *systèmes de recommandation d'articles de recherche* et les moteurs de recherche spécialisés aident les chercheurs à localiser de nouvelles ressources, mais là encore échouent dans l'aide à les gérer. Finalement, les *systèmes de Gestion de Contenu D'entreprise* offrent les fonctionnalités de gestion de documents et des connaissances, mais ne sont pas conçus pour les articles de recherche. Dans ce mémoire, nous présentons une nouvelle classe de systèmes de gestion : *système de gestion et de recommandation d'articles de recherche*. *Papyres* (Naak, Hage, & Aïmeur, 2008, 2009) est un prototype qui l'illustre. Il combine des fonctionnalités de bibliographie avec des techniques de recommandation d'articles et des outils de gestion de contenu, afin de fournir un ensemble de fonctionnalités pour localiser les articles de recherche, manipuler et maintenir les bibliographies. De plus, il permet de gérer et partager les connaissances relatives à la littérature. La technique de recommandation utilisée dans Papyres est originale. Sa particularité réside dans l'aspect *multicritère* introduit dans le processus de *filtrage collaboratif,* permettant ainsi aux chercheurs d'indiquer leur intérêt pour des *parties* spécifiques des articles. De plus, nous proposons de tester et de comparer plusieurs approches afin de déterminer le voisinage dans le processus de *Filtrage Collaboratif Multicritère*, de telle sorte à accroître la précision de la recommandation. Enfin, nous ferons un rapport global sur la mise en œuvre et la validation de Papyres.

**Mots-clés :** Gestion d'Articles de Recherche, Gestion de références, Gestion de Contenu d'Entreprise, Système de Recommandation d'Articles de Recherche, Filtrage Collaboratif Multicritère, classification des Systèmes de Recommandation.

# Abstract

Graduate students and professors (researchers, in general) regularly access, review, and use large amounts of research papers, yet none of the existing tools and solutions provides the wide range of functionalities required to properly manage these resources. Indeed, *bibliography management systems* manage the references and citations but fail to help researchers in handling and locating resources. On the other hand, *research paper recommendation systems* and specialized search engines help researchers to locate new resources, but again fail to help researchers in managing the resources. Finally, *Enterprise Content Management systems* offer the required functionalities to manage resources and knowledge, but are not designed for research literature. Consequently, we suggest a new class of management systems: *Research Paper Management and Recommendation System*. Through our system *Papyres* (Naak, Hage, & Aïmeur, 2008, 2009) we illustrate our approach, which combines bibliography functionalities along with recommendation techniques and content management tools, in order to provide a set of functionalities to locate research papers, handle and maintain the bibliographies, and to manage and share knowledge related to the research literature. Additionally, we propose a novel research paper recommendation technique, used within Papyres. Its uniqueness lies in the *multicriteria* aspect introduced in the process of *collaborative filtering*, allowing researchers to indicate their interest in specific *parts* of articles. Moreover, we suggest test and compare several approaches to determine the neighbourhood in the *Multicriteria Collaborative Filtering* process, such as to increase the accuracy of the recommendation. Finally, we report on the implementation and validation of Papyres.

**Keywords :** Research Paper Management, Reference Management, Enterprise Content Management, Research Paper Recommendation, Multicriteria Collaborative Filtering, Recommendation Systems' classification.

# Table des matières

# Liste des tableaux

# Liste des figures

*à*

*mes parents et ma famille*

# Remerciements

*Mes sincères remerciements et ma profonde reconnaissance vont à ma directrice de recherche, Professeure Esma Aïmeur, pour m'avoir dirigé et soutenu tout au long de ce projet, surtout dans les moments difficiles. Votre esprit scientifique et votre souci pour de hautes performances sont une grande inspiration et m'ont poussé à me surpasser. Merci pour vos efforts, vos conseils et vos critiques constructives. Merci pour tout !*

*Des remerciements particuliers vont à Hicham Hage en signe de reconnaissance pour son implication, ses efforts et sa disponibilité. Les nombreuses discussions que nous avions eues et ses suggestions fructueuses ont bien marqué ce mémoire.*

*Je tiens aussi à remercier tous mes ami(e)s qui m'ont aidé de loin ou de près pour l'accomplissement de cet ouvrage, en particulier, ceux qui m'ont offert de leur précieux temps pour lire et réviser ce mémoire.*

*Je remercie également mes collègues de notre laboratoire Héron qui ont su créer une belle et chaleureuse ambiance et une expérience académique des plus riches dans un climat où règne le respect mutuel.*

*J'adresse également mes remerciements aux membres du jury qui ont accepté d'être rapporteurs de mon mémoire.*

*À mes chers parents et ma famille, je dédie ce modeste travail.*

*Je garde le mot de la fin pour ma femme Naima et mes enfants, Lyna et Samy, je les remercie profondément de m'avoir pardonné mon indisponibilité et pour m'avoir encouragé et soutenu durant cette période.*

*Que chacun(e) trouve ici l'expression de ma grande gratitude et sympathie !*

*Amine Naak*

# Chapitre 1 Introduction

Le monde de la science sait pertinemment le rôle que s'octroie un *article de recherche* dans la conquête de l'univers du savoir. Un univers en permanente activité, qui ne cesse de pousser ses frontières de jour en jour avec la publication de travaux dans les *conférences* et les *journaux scientifique*. La migration, des supports traditionnels vers les formats numériques avec l'avènement des nouvelles technologies de l'information et de la communication (NTIC), a multiplié la vitesse de production de ces articles, et a permis la naissance de *Bibliothèques numériques*, comme IEEExplore (URL, 1), ACM *digital librairies* (URL, 2) et SpringerLink (URL, 5) dont la base de données avoisine les sept millions d'articles tous domaines confondus (Reuters, 2008). Cette technologie a bouleversé les repères spatio-temporels et économiques de notre vie. Depuis, ni le transfert d'un bout du monde à l'autre ni la production d'une quasi-infinité de copies de fichiers n'a de coût significatif en temps et en argent.

D'une part, cette migration du monde traditionnel vers l'internet, a rendu l'accès à l'information, plus facile et plus rapide comme jamais auparavant. À la manière de la collection d'images, elle peut varier d'un simple album numérique à une forme de réseau social de partage, les chercheurs scientifiques peuvent entretenir leurs propres bibliothèques numériques ou se les partager avec un groupe de participants, indépendamment de leur localisation. Ce partage s'étend aux méta-informations de tout genre, telles que les commentaires, les résumés, les revues, les évaluations, les étiquettes (*Tags)*. D'autre part, ce beau monde a engendré d'autres besoins et soucis pour les chercheurs qui se voient imposer une nouvelle cadence et un nouveau rythme qu'ils doivent adopter pour survivre dans cette nouvelle ère. Ce cumul d'informations durant le cycle de vie des articles, combiné à leur grand nombre et la rapidité à laquelle ils se produisent, a engendré un dépassement de nos facultés intellectuelles.

Les anciens défis sont amplifiés et des nouveaux s'y ajoutent. Désormais, pour atteindre ses objectifs, le chercheur mènera sa bataille sur différents fronts. Conséquence d'une haute production d'informations, la tâche du chercheur est de plus en plus compliquée, notamment pour comprendre et rester à jour avec les nouveautés de son domaine. Il est primordial de bien connaitre son domaine et les problèmes qui y sont liés, ne dit-on pas que comprendre la question, est la moitié de la réponse ? En effet, avant de résoudre un problème, il faut bien le comprendre, et cela passe inévitablement par une vaste documentation et une lecture des ressources[1] dans le domaine. Le partage des différentes notes sur les articles dans une sorte de réseau social où les chercheurs rendent publics leurs critiques, leurs commentaires, leurs résumés, etc. est d'une très grande utilité. Il facilite la compréhension et permet un gain de temps non négligeable. Pour se mettre à jour, le chercheur doit enrichir continuellement son état de l'art en étendant sa base d'articles. Il doit aussi maintenir une veille technologique sur les nouveautés qui peuvent surgir d'un instant à l'autre, en provenance de divers horizons, jusqu'à la publication de ses résultats. La revendication d'un résultat en se prévalant des droits d'auteur exige de la part du chercheur de s'assurer de l'originalité de ses travaux, bien avant leur publication. Autrement, il risque d'ignorer des travaux similaires et de ce fait, violer leurs droits.

La négligence, de l'organisation et de la gestion de ces articles, mène à un désarroi et à la confusion, ce qui pénalise le chercheur en question par une perte de temps, d'efforts, d'informations voir même d'argent. Nul doute, pour pénétrer et laisser son empreinte, les chercheurs doivent optimiser leur effort et leur temps en les coordonnant avec des moyens technologiques adéquats à la hauteur de ces défis. Des moyens qui leur faciliteront la gestion et la localisation de ces ressources.

L'objectif de ce mémoire est d'étudier en détail ces problèmes et de recenser les différents besoins dans le domaine. Par la suite, nous allons faire le constat des solutions susceptibles de répondre à ces besoins ou du moins nous en inspirer pour produire un prototype illustrant cette nouvelle classe de systèmes.

---

[1] Dans ce travail, nous nous référons aux divers genres de littérature de recherches (article de conférence, article de journal, livres, rapports, etc.) par le terme « ressource ».

Une première solution, à l'organisation et la gestion d'articles de recherche, consiste à utiliser *le système de gestion de fichiers* fourni comme outil de base avec tout système d'exploitation. Donc, les articles seront classés dans une arborescence de dossiers et sous-dossiers dans un média comme un disque dur. Cette solution n'est pas adéquate pour diverses raisons, entre autres la classification multiple d'un article par rapport à ses multiples attributs causera à court terme une grande redondance, un désordre et une difficulté de repérage. La complexité croît avec le nombre d'articles et de critères de classement. Une deuxième solution est l'utilisation des *systèmes de gestion de base de données* (SGBD). L'article est manipulé indépendamment de sa copie numérique, désormais il est représenté par un sous-ensemble de ses métadonnées désigné par le terme *référence* ou *citation*. L'ensemble de ces métadonnées est sauvegardé dans des bases de données faciles à manipuler ainsi qu'un lien optionnel vers la copie numérique du document accessible grâce à l'utilisation du système de gestion de fichiers. Cette solution est bien adaptée aux types de données structurées. Mais, elle est peu efficace lorsqu'il s'agit de contenus non structurés, comme les images, les vidéos et les contenus textes, qui sont parmi les types considérés dans le cadre de ce mémoire. Une troisième solution consiste à utiliser des techniques de gestion de contenu, précisément, celles utilisées dans les ECM (*Entreprise Content Management*). Ces systèmes s'appuient sur les solutions précédentes et se servent de plusieurs autres techniques spécialisées dans la gestion de contenus non structurés. Par exemple, elles se servent des techniques WCM (*Web Content Management*) pour séparer les données de leur mise en forme pour faciliter leur publication. Les techniques de recherche d'informations sont un autre type de manipulation de contenu qui permet d'explorer le texte du document.

Sur le plan produit technologique, il existe divers logiciels ou applications qui implémentent certaines des solutions précédentes. Ces applications, avec des niveaux de complexité différents, varient d'un stade prototype, à des solutions plus complètes, gratuites ou commerciales. Certaines de celles-ci sont spécialement dédiées pour les articles de recherche, alors que d'autres sont conçus pour un contexte différent, mais elles présentent beaucoup de similitudes par rapport à l'objet de notre recherche. Ce mémoire présente une étude détaillée de ces applications et il les répartit suivant des catégories distinctes. Chaque

catégorie est illustrée par des études de cas représentatifs de ses sous-classes afin de montrer leurs points forts et leurs faiblesses par rapport à nos besoins. Toutes ces applications rentrent dans les trois classes suivantes : systèmes de gestion bibliographique, systèmes de gestion de contenu et systèmes de recommandation.

## Les systèmes de gestion bibliographique

L'article de recherche est le moyen principal de publication et d'information pour les chercheurs scientifiques, tous domaines confondus. Un chercheur voulant annoncer les résultats de ses travaux, le fera avec la publication d'articles dans les conférences et les journaux scientifiques correspondants. De même, pour s'informer des dernières nouveautés de son domaine, il consultera ces mêmes ressources. D'où la nécessité d'avoir des applications spécialisées et centrées autour de ces articles de recherche. Les plus spécialisées sont les *systèmes de gestion de références* qu'on trouve implémentés suivant deux architectures différentes : application traditionnelle basée en local (ordinateur personnel) et application basée sur le Web. Un exemple de celles basées en local est EndNote (URL, 3); il permet de décharger les chercheurs des tâches consistant à organiser et à formater les références suivant plusieurs *styles bibliographiques* (IEEE, APA, etc.). Avec son module qui s'intègre facilement dans un logiciel de traitement de texte, EndNote permet de gérer automatiquement la bibliographie en fin de document et d'ajouter la référence au bon endroit en respectant l'ordre prédéfini. Le point fort de ces applications est justement cette manipulation des références, mais elles ont failli dans tout ce qui a trait aux fonctionnalités de nature réseau social. Les applications basées sur le Web rattrapent ce manque, c'est justement leur point fort. Un exemple de celles-ci : CiteUlike(URL, 4) et les outils BibTex[2] (URL, 6) tels que Bibshare[3]; ils permettent aux chercheurs de saisir diverses informations sur les articles, notamment, des commentaires, des étiquettes (*tags*), des revues. Ces applications sont capables de gérer séparément les espaces privés, de groupes ou publics. Elles permettent aux usagers de grouper des articles[4] dans des bibliothèques

[2] Le terme BibText désigne à la fois un format de fichier et les applications qui le manipulent.
[3] http://bibshare.dsic.upv.es/
[4] Le terme article de recherche ou brièvement article est employé indifféremment pour désigner sa référence.

privées, partagées au sein d'un groupe ou de les rendre publics, accessibles pour tout le monde. Sur le plan localisation de ressources ou gestion de documents, et quel que soit son architecture, cette classe de système, présente différentes implémentations élémentaires qui sont loin de satisfaire les exigences des chercheurs en cette matière.

## Les systèmes de recommandation d'articles de recherche

Les systèmes de gestion de bibliographie ne fournissent pas assez de fonctionnalités importantes nécessaires pour les chercheurs qui sont régulièrement à la conquête de nouvelles ressources. Si l'on considère la grande masse d'articles disponibles dans les bibliothèques numériques et leur rapidité de production et diffusion, la localisation d'articles est une tâche fastidieuse, malgré l'utilisation de moteurs de recherche tels que Google Scholar (URL, 7) ou CiteSeer (URL, 8). Les résultats retournés demeurent superflus du fait qu'ils se basent uniquement sur des mots clés et ne tiennent pas compte d'informations de profil. Cela nous amène à considérer une autre classe de systèmes qui tient compte de ce manque, et ce sont les systèmes de recommandation d'articles de recherches. Contrairement au moteur de recherche, ces derniers retournent des résultats personnalisés parmi un grand espace d'informations en se basant sur diverses considérations, comme dans le cas du système TechLens (Kapoor et al., 2007). Celui-ci emploie les listes de références des utilisateurs pour établir un profil sur lequel les recommandations seront basées. Néanmoins, ces systèmes manquent de fonctionnalités concernant la production et la gestion des méta-informations relatives à l'article de recherche. Un chercheur qui est en train de lire un article, souhaiterait par exemple le commenter, le critiquer. Il pourrait aussi avoir besoin de les classer dans des dossiers et sous-dossiers; pouvoir effectuer le suivi de ses documents à l'aide de fonctionnalités qui le renseignent rapidement s'il est lu ou pas, commenté ou pas, etc. Avec un tel manque, les chercheurs ne peuvent pas facilement garder une trace de leurs critiques et analyses. Il est important d'offrir des moyens pour faciliter la gestion de telles informations, car elles reflètent des parties de la connaissance *implicite* du chercheur; une connaissance qu'on ne trouvera nulle part ailleurs et qui fait appel au savoir et à l'expérience du chercheur.

Certaines de ces applications offrent quelques fonctionnalités de gestion, mais elles ne répondent pas aux besoins de base des chercheurs.

## Les systèmes de gestion d'entreprise

Les systèmes **ECM** (***Enterprise Content Management***), tels que le LiveLink (URL, 10), offrent des outils et des fonctionnalités pour contrôler une telle connaissance implicite. Néanmoins, ces outils sont conçus pour un environnement corporatif, et sont créés spécifiquement pour la gestion interne des documents, des projets, le *workfow*, etc. Ainsi, bien que ces systèmes offrent beaucoup de fonctionnalités, une grande part d'entre elles ne sont pas directement liées à la recherche, et ne prennent pas en considération les spécificités des articles de recherche. L'architecture de ces systèmes est complexe; elle est composée de plusieurs modules implémentant de nombreuses fonctionnalités. Plusieurs de celles-ci, comme les fonctionnalités d'archivage, ne sont pas pertinentes pour l'objet de notre recherche. Toutefois, parmi les fonctionnalités qui nous intéressent, certaines ne sont pas directement applicables dans notre contexte; elles nécessitent des adaptations afin de prendre en considération les particularités de l'article de recherche. Malgré cette complexité, il y a toute une classe de fonctionnalités qui n'est pas implémentée, en l'occurrence celle spécialisée dans la gestion de références.

Comme il apparait après cette courte analyse, les trois classes de systèmes précédents ne répondent pas aux besoins de fonctionnalités pour la recherche; leur rassemblement dans un seul système constitue une nouvelle classe; c'est l'une des contributions de ce mémoire.

## Originalité et contribution

Ce mémoire présente une contribution au domaine de la gestion et de la recommandation d'articles de recherche avec l'implémentation de *Papyres* (Naak, Hage, & Aïmeur, 2008, 2009). Cette contribution vue selon quatre points : une nouvelle classe de systèmes de gestion de contenu, introduction d'une nouvelle vision pour l'article de recherche, application du filtrage collaboratif multicritère dans ce domaine, et finalement l'optimisation de l'algorithme de filtrage collaboratif multicritère.

- **Une nouvelle classe de systèmes de gestion de contenu :** Ce mémoire décrit une nouvelle classe de systèmes, en l'occurrence le système de gestion d'articles de recherche dont un prototype a été implémenté et nous l'avons appelé Papyres. Sa première originalité est le fait qu'il combine les fonctionnalités de bibliographie avec les outils de gestion de contenu (ECM) et les techniques de recommandation, afin de fournir un ensemble de fonctionnalités pour gérer les références, pour organiser, gérer et localiser les travaux de recherche, et pour partager et gérer les différentes méta-informations. D'ailleurs, Papyres favorise le Web2.0 (O'Reilly, 2005), et utilise des techniques Web2.0 et des technologies telles que l'étiquetage (*Tagging*), l'évaluation (*Rating*), et le RSS (*Really Simple Syndication*).
- **Une nouvelle vision pour l'article de recherche :** Notre application exploite une vue originale de l'article de recherche, qui consiste en la segmentation de ces articles en parties prédéfinies : Introduction, État de l'art, Expérimentation, etc. Le système offre la possibilité de les évaluer, les commenter, et les *recommander séparément* et indépendamment de l'article global.
- **Application du filtrage collaboratif multicritère dans ce domaine :** à notre connaissance, Papyres constitue la première application qui utilise un filtrage collaboratif multicritère dans la recommandation d'articles de recherche. On trouve, dans la littérature publiée, certains travaux qui parlent de la recommandation multicritère, mais généralement, différents de notre approche. En effet, ils sont basés sur les méthodes issues de la recherche opérationnelle. L'article (Manouselis & Costopoulou, 2007) constitue un exemple de cela.
- **Optimisation algorithmique du filtrage collaboratif multicritère :** Nous avons proposé plusieurs alternatives à ceux proposés dans le seul article (Adomavicius & Kwon, 2007). Nous avons effectué plusieurs expériences et nous avons comparé les résultats. Nous avons pu constater la distinction de la méthode HZ-N (Horizontale sans Bruit) par rapport aux autres y compris celles proposées par l'article précédent.

## Organisation de ce mémoire

Ce mémoire est organisé comme suit : le chapitre 2 offre un survol des systèmes de recommandation. Ce survol peut être divisé en deux parties. Dans la première nous discutons plusieurs logiques de classification bien connues dans ce domaine pour mettre en relief la classe à laquelle s'apparente notre méthode. Dans la deuxième partie, nous introduisons la recommandation multicritère, plus précisément le filtrage collaboratif comme préalable à la présentation de notre approche multicritère que nous avons appliquée à la recommandation d'articles de recherche. Le chapitre 3 présente et discute des solutions susceptibles de répondre aux besoins de la recherche préalablement analysés. Nous montrons comment ces systèmes ne répondent pas à ces besoins. Ce chapitre représente un état de l'art des systèmes suivants : les systèmes de gestion de références, les systèmes de gestion de contenu d'entreprise et finalement les systèmes de localisation et de recommandation d'articles de recherche. Le chapitre 4 présente Papyres, notre solution pour combler les besoins présentés dans la section précédente. Nous décrivons le détail des fonctionnalités implémentées. Le chapitre 5 présente les résultats de la validation globale de notre système, ainsi que les résultats de tests et comparaisons des différentes approches utilisées dans le calcul du voisinage. Tout cela après avoir présenté les aspects technologiques entourant l'implémentation de ce système ainsi que la description des différentes interfaces de l'application. Et finalement, le dernier chapitre conclut ce mémoire et présente quelques perspectives futures.

# Chapitre 2 Les systèmes de recommandation

L'objectif de ce chapitre est de faire un bref survol du domaine des systèmes de recommandation à travers lequel nous abordons leurs classifications et la terminologie utilisée. L'approche, que nous suivons, est celle de confronter certaines vues très répondues dans ce domaine et tenter de les rapprocher. Nous présentons aussi, la recommandation multicritère qui est une nouvelle tendance dans ce domaine et comme un préalable nécessaire pour comprendre notre apport dans ce mémoire. À la fin, notre conclusion détermine la logique et la terminologie adoptée tout au long de ce mémoire.

## 2.1 Introduction

Les systèmes de recommandation peuvent être considérés comme une sorte d'outils, pour qui un usager divulgue son profil (représenté, d'une façon générale, par ses préférences et ses goûts), en contrepartie, ils lui projettent une image personnalisée et réduite de toute une masse d'informations qui dépasse nos facultés cognitives. À première vue, ils ressemblent aux moteurs de recherches d'informations (Burke, 2002), d'ailleurs ils peuvent être considérés comme de simples systèmes de recommandations qui se contentent de retourner une liste de résultats ordonnés suivant un élément du profil de cet usager, lequel dans ces systèmes, se réduit, tout simplement, aux mots clés de sa requête. De ce fait, les résultats retournés, pour une même liste de mots clés, sont identiques, pourvu que la requête soit la même. En l'absence de systèmes de recommandation, un utilisateur, en quête de l'information sur le Web et face à cette marée d'informations, a recours à ces moteurs de recherche, parmi lesquels se trouvent Google, Yahoo!, Altavista. La nature de l'information manipulée par ces moteurs de recherche est textuelle sous de multiples formats par exemple le format HTML des pages Web. Le domaine d'application des systèmes de recommandation peut porter sur un très vaste espace d'items comme : les

films, les musiques, les images, les produits commerciaux. D'ailleurs, ils font de plus en plus partie intégrante des sites de commerce électronique.

### 2.1.1 Histoire et définitions

Les premiers systèmes de recommandation se réduisent aux systèmes de filtrage collaboratif. Ils remontent au début des années 1990s, cette période à laquelle ils sont reconnus comme étant un domaine de recherche indépendant. Parmi les systèmes pionniers dans ce domaine, nous citons à titre d'exemple les systèmes : Tapestry (Goldberg, Nichols, Oki, & Terry, 1992), GroupLens/NetPerceptions (Resnick, Iacovou, Suchak, Bergstrom, & Riedl, 1994), Ringo/Firefly (Shardanand & Maes, 1995). Les racines des systèmes de recommandation remontent aux travaux étendus dans les sciences cognitives, la théorie d'approximation, la recherche documentaire, la théorie de la prévoyance et ont également des liens avec la science de la gestion et le marketing, dans la modélisation des choix du consommateur (Adomavicius & Tuzhilin, 2005).

Intuitivement, un système de recommandation, par exemple de livres, peut être comparé à un agent de bibliothèque à qui on demande de proposer des ouvrages après lui avoir exprimé nos besoins. En effet, le bibliothécaire avec sa vaste connaissance des ouvrages de cette bibliothèque comparativement à un simple usager et compte tenu de sa longue expérience avec les usagers ainsi qu'un feedback en provenance de ces derniers peut prodiguer de précieuses recommandations qui épargneront un dur effort de recherche et beaucoup de temps en conséquence. Bien sûr, l'usager doit lui fournir suffisamment de données. Cet exemple illustre un *recommander* humain, afin de rapprocher l'image à l'échelle de la machine, le bibliothécaire virtuel doit avoir trois sortes d'informations :

- La représentation de la bibliothèque (profil des items)
- La représentation des usagers (profil des usagers)
- Les algorithmes qui interfèrent entre les deux pour produire les recommandations

En d'autres termes, un système de recommandations aide les utilisateurs à faire leurs choix dans un domaine où ils disposent de peu d'informations pour trier et évaluer les alternatives possibles (Resnick & Varian, 1997; Shardanand & Maes, 1995). L'importance

de tels systèmes apparaît clairement dans les environnements où de gigantesques quantités d'informations en ligne surpassent de loin les capacités de recherche de n'importe quel être humain. Actuellement, les systèmes de recommandation font partie intégrante de certains sites de commerce électronique tels qu'Amazon[1] et CDNow[2].

## 2.2 Classification des systèmes de recommandation

Depuis l'émergence du domaine des systèmes de recommandation, plusieurs approches et méthodes ont été proposées, certaines ont été étudiées, expérimentées et comparées. Parfois plusieurs terminologies sont utilisées pour désigner une même méthode ou approche. C'est pour cela que certains chercheurs se sont intéressés à la classification de ces méthodes et proposent une taxonomie ou terminologie unifiée. Nous citons les travaux de [Adomavicius & Tuzhilin 2005; Burke 2002; Manouselis & Costopoulou 2007; Miquel, Beatriz *et al.* 2003; Pazzani 1999; Resnick & Varian 1997; Schafer, Konstan *et al.* 1999]. Particulièrement, les deux articles (Adomavicius & Tuzhilin, 2005; Burke, 2002) ont fait un survol très intéressant pour rassembler les différents points de vues. D'ailleurs, de plus en plus de travaux se basent sur ces derniers.

Une autre étude, en l'occurrence (Manouselis & Costopoulou, 2007), en plus de la recommandation traditionnelle, se distingue par la classification d'une tout autre catégorie de systèmes de recommandation dite *systèmes de recommandation multicritère* qui est basée sur les méthodes issues des sciences d'aide à la décision, MCDMs (*Multi-Criteria Decision Making*). Également abordé dans (Adomavicius & Tuzhilin, 2005), les auteurs se sont contentés de donner une idée générale à titre d'extension possible des systèmes traditionnels et ce n'est que récemment dans l'article (Adomavicius & Kwon, 2007) les auteurs ont proposé une extension des méthodes traditionnelles, dites monocritères, pour considérer le cas du multicritère.

La classification proposée, dans ce qui suit, ne tient pas compte des travaux de (Manouselis & Costopoulou, 2007), du fait que notre système multicritère (Naak, Hage, &

[1] www.amazon.com
[2] www.cdnow.com

Aïmeur, 2008, 2009) s'inscrit dans la même perspective que (Adomavicius & Kwon, 2007; Adomavicius & Tuzhilin, 2005).

Le survol de la littérature des systèmes de recommandation, entre autre (Adomavicius & Tuzhilin, 2005; Balabanovi & Shoham, 1997; Burke, 2002; Resnick & Varian, 1997; Schafer, Konstan, & Riedl, 1999), montre qu'il y a un consensus en ce qui concerne la classification des systèmes de recommandation pour les trois catégories : les méthodes basées sur *le filtrage collaboratif*, les méthodes *basées sur le contenu* et les méthodes *Hybrides*. Néanmoins, cette dernière classe ne représente pas une méthode spécifique, mais toute autre méthode différente des deux premières, y compris celles résultantes de leur combinaison. La divergence majeure réside justement dans la classification de ces hybrides. La suite de ce chapitre est basée sur les deux classifications citées ci-dessus : (Adomavicius & Tuzhilin, 2005; Burke, 2002) malgré leurs points de vue différents, ils se ressemblent beaucoup. Après la présentation de leur logique de classification, nous commencerons par détailler les méthodes classiques. Après une discussion de leurs faiblesses, nous présentons les méthodes hybrides et nous terminons par un résumé qui tente de rapprocher les deux approches.

Dans ce mémoire, les termes *classique* et *pur*, dans le contexte de la recommandation, font référence aux méthodes basées sur le filtrage collaboratif et les méthodes basées sur le contenu. Elles sont classiques du fait qu'elles sont les premières utilisées et qu'elles sont largement admises. Elles sont pures par comparaison aux hybrides, qui justement sont des combinaisons de ces deux classiques.

### 2.2.1 Approche de classification de Burke (2002)

"*Of interest in this discussion is not the type of interface or the properties of the user's interaction with the recommender, but rather the sources of data on which recommendation is based and the use to which that data is put.*"(Burke, 2002).

Pour Burke tout système de recommandation est constitué de trois éléments :

- Les données préalables : ce sont les informations que le système possède avant le processus de recommandation.

- Les données d'entrée : ce sont les informations que l'usager doit communiquer au système dans le but de lui générer une recommandation
- Un algorithme qui combine les données préalables et les données d'entrée pour parvenir à ses suggestions.

En général, cette classification introduit trois autres méthodes par rapport à celles communément admises, notamment, le *filtrage démographique*, le *filtrage à base d'utilité* et le *filtrage à base de connaissance*. Même s'il distingue le filtrage démographique comme une méthode à part entière, Burke, fait remarquer qu'elle est un cas particulier du filtrage collaboratif. La même chose pour la méthode à base d'utilité, celle-ci peut être associée à la méthode à base de connaissance. Donc, finalement ce qui reste comme plus est le filtrage à base de connaissances.

### 2.2.2 Approche de classification de Adomavicius et Tuzhilin (2005)

Une des originalités de cette approche réside dans le fait que les auteurs ont proposé une formulation du problème de la recommandation comme suit :

$$\forall c \in C, \qquad s'_c = \arg\max_{s \in S} u(c, s)$$

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

$C$ : ensemble représentant tous les usagers

$S$ : ensemble de tous les items qui sont recommandables

$u$ : une fonction d'utilité qui mesure l'utilité de l'item $s$ pour l'usager $c$

$R$ : ensemble totalement ordonné

Le problème de la recommandation revient à trouver pour chaque usager $c \in C$, un certain item $s' \in S$, qui va maximiser la fonction d'utilité $u : C \times S \rightarrow R$. Dans les systèmes de recommandation, l'utilité d'un item est exprimée par une évaluation sur une échelle de valeurs définie qui représente l'ensemble $R$.

Le centre du problème des systèmes de recommandation est la fonction d'utilité $u : C \times S \rightarrow R$. Généralement, elle n'est pas définie complètement sur tout l'ensemble $C \times S$, mais uniquement sur un sous ensemble de ce dernier. Typiquement la fonction d'utilité dans les systèmes de recommandation est représentée par les évaluations données par les usagers. La recommandation consiste en l'extrapolation de ces évaluations sur tout l'espace, en prédisant les valeurs inconnues. Les prédictions qui obtiennent les meilleurs scores feront l'objet de la recommandation.

Inspirés de plusieurs travaux dans le domaine des systèmes de recommandation, leurs logiques reposent sur trois concepts :

**1) Types d'algorithme de recommandation :**

L'extrapolation à partir d'évaluations connues vers celles qui ne le sont pas est, généralement, réalisée de deux façons : par la spécification d'une *heuristique* ou par l'utilisation d'un *modèle* (Balabanovi & Shoham, 1997). La technique basée sur une *heuristique* dite aussi basée sur la *mémoire* consiste à utiliser les évaluations des usagers à chaque fois qu'une recommandation est calculée. Par contre, la technique basée sur un modèle n'utilise pas directement les évaluations pour générer les recommandations comme la précédente, mais elle les utilise pour apprendre un modèle, pour ensuite l'utiliser afin de prodiguer des recommandations.

**2) Le profil usager et le profil item :**

Chaque usager de l'espace *C* est représenté par un profil qui est constitué de plusieurs informations ou caractéristiques, telles que l'âge, le sexe, les revenus, son état matrimonial. Dans le plus simple des cas, le profil contient uniquement son identifiant. De même, pour chaque item de l'espace *S* des items, on peut lui associer un profil. Par exemple, un livre peut être représenté par son titre, son auteur, sa langue, son genre.

**3) Symétrie des algorithmes par rapport à l'usager ou à l'item :**

Il y a une symétrie quant à la recommandation par rapport à l'usager ou à l'item. En effet, on peut alternativement recommander les *N* meilleurs items à un certain usager, comme on peut trouver les *N* meilleurs usagers pour un item donné. Ainsi toutes les méthodes qui s'appliquant à un cas, s'appliquent à l'autre.

## 2.3 Les méthodes classiques et pures

### 2.3.1 Le filtrage collaboratif et le filtrage démographique

C'est la méthode qui était à l'origine des systèmes de recommandation et c'est la plus appliquée de toutes les autres. Désignée aussi par le terme « *Word of Mouth* » (Shardanand & Maes, 1995) ou encore par «*people to people correlation*» (Schafer, Konstan, & Riedl, 1999), mais le terme filtrage collaboratif demeure le plus populaire. Intuitivement, un système basé sur le filtrage collaboratif recommande à un usager donné, les items hautement évalués par d'autres usagers qui présentent des similarités dans leurs goûts et préférences.

D'après (Breese, Heckerman, & Kadie, 1998), sur le plan algorithmique, il y a deux classes de filtrage collaboratif : les algorithmes basés sur la mémoire (*memory-based*) dits aussi basés sur les heuristiques (*heuristic-based*) et ceux basés sur les modèles (*model-based*).

#### 2.3.1.1 Filtrage collaboratif basé sur la mémoire ou Heuristique

Le filtrage collaboratif à base de mémoire ou heuristique considère la totalité des évaluations des usagers disponibles au moment du calcul de la recommandation. Le processus de calcul de la recommandation pour un usager $u_i$ passe par deux étapes successives :

- Phase du calcul du voisinage

En se basant sur le profil de cet usager $u_i$, le système recherche les usagers $u_j$ ($j$ diffère de $i$) qui lui sont les plus similaires. Deux mesures de similarité qui sont très

utilisées sont : *la similarité vectorielle* et *la corrélation de Pearson* (Breese, Heckerman, & Kadie, 1998)

- La similarité vectorielle

Dans cette méthode les usagers *A* et *B* sont considérés comme deux vecteurs de même origine dans un espace de *m* dimensions, *m* est égale au nombre d'items évalués par les deux usagers. Plus deux usagers sont similaires, plus l'angle entre leur vecteur est plus petit. Empiriquement, la similarité entre ces deux usagers est calculée par la formule du *Cosinus* suivante :

$$\cos inus(A,B) = \sum_{j=1}^{l1} \frac{v_{A,j}}{\sqrt{\sum_{j=1}^{l1} v_{A,j}^2}} \times \frac{v_{B,j}}{\sqrt{\sum_{j=1}^{l1} v_{B,j}^2}}$$

*l1* : nombre d'items communs entre *A* et *B* votés par *v*.

$v_{A,j}$ : vote de *A* pour l'item *j*

$v_{B,j}$ : vote de *B* pour l'item *j*

**Formule 4.1 :** Le calcul du cosinus

- La corrélation de Pearson

La corrélation de Pearson telle qu'utilisée par (Resnick, Iacovou, Suchak, Bergstrom, & Riedl, 1994; Shardanand & Maes, 1995) est une méthode issue des statistiques. Elle est aussi très utilisée dans le domaine des systèmes de recommandation pour mesurer la similarité entre deux usagers. La formule ci-dessous, nous donne cette valeur pour deux usagers *A* et *B* :

$$w(A,B) = \frac{\sum_j (v_{A,j} - \overline{v_A})(v_{B,j} - \overline{v_B})}{\sqrt{\sum_j (v_{A,j} - \overline{v_A})^2 \sum_j (v_{B,j} - \overline{v_B})^2}}$$

*j :* nombre d'objets ayant été voté à la fois par *A* et *B*

$v_{A,j}$ : vote de *A* pour l'item *j*.

$\overline{v_A}$ : moyenne des votes de *A*.

**Formule 2 :** La corrélation de Pearson

- Phase de prédiction

Une fois que toutes les similarités de l'usager cible *A* par rapport aux autres usagers sont calculées et que les *n* usagers les plus similaires qui constituent le voisinage de cet usager cible sont définis, la prédiction de la valeur d'un item *j* évaluée par l'utilisateur *A* est calculée à l'aide de la formule suivante :

$$p_{A,j} = \overline{v_A} + \frac{\sum_{i=1}^{n} w(A,i)(v_{i,j} - \overline{v_i})}{\sum_{i=1}^{n} |w(A,i)|}$$

*n :* nombre d'utilisateurs présents dans le voisinage de *A,* ayant déjà voté sur l'objet *j*.

$v_{i,j}$ : vote de l'utilisateur *i* pour l'objet *j*.

$\overline{v_i}$ : moyenne des votes de l'utilisateur *i*.

**Formule 3 :** Formule de calcul de la prédiction

#### 2.3.1.2 Filtrage collaboratif basé sur le modèle

Les algorithmes basés sur le modèle se basent aussi sur les évaluations précédentes (les profils) des usagers, sauf que cette fois-ci, on ne calcule pas directement les prédictions, mais on essaie de classifier les usagers suivant des groupes ou d'apprendre les modèles à partir de leurs données. Une fois les groupes ou les modèles d'usagers sont trouvés, la prédiction pour un usager donné est générée automatiquement à partir de son

profil. Pour la construction du modèle plusieurs méthodes sont utilisées, par exemple (Breese, Heckerman, & Kadie, 1998) qui proposent deux méthodes statistiques pour construire les modèles, la première est très utilisée dans le domaine de l'apprentissage machine : Les modèles de classification (en Anglais *Cluster models*). La deuxième méthode est celle des réseaux Bayesians. En général, les méthodes basées sur le modèle utilisent les techniques d'apprentissage machine et les techniques statistiques pour apprendre le modèle à partir des profils des usagers.

### 2.3.2 Le filtrage à base de contenu

Les méthodes basées sur le contenu, comme leur nom l'indique, se basent sur la compréhension de pourquoi l'usager, à qui la recommandation est destinée, a donné une haute valeur à certains items qu'il a évalués dans le passé ? Une fois cette question résolue, le système cherche parmi les nouveaux items ceux qui maximisent ces caractéristiques pour les lui recommander. Ou comme (Pazzani, 1999) l'a défini : « *Content-based methods make recommendations by analyzing the description of the items that have been rated by the user and the description of items to be recommended* » dont la traduction est : les méthodes basées sur le contenu émettent des recommandations en analysant la description des articles qui ont été évalués par l'utilisateur et la description des articles destinés à être recommandés.

Habituellement, cette méthode est utilisée dans la recommandation d'items de type textuel, comme les articles de journaux, les pages Web. D'ailleurs, cette méthode s'appuie sur les techniques de recherche d'informations (Baeza-Yates & Ribeiro-Neto, 1999; Salton, 1989) ou les techniques de filtrage d'informations (Belkin & Croft, 1992) qui ont réalisé une grande avancée dans le domaine de la recherche basée sur le contenu. Le principal apport des systèmes de recommandation est la prise en considération des profils qui représentent les goûts et les préférences de l'usager. Le profil de l'usager peut être récolté directement à travers par exemple un questionnaire ou d'une façon implicite à travers son comportement au fil du temps (Adomavicius & Tuzhilin, 2005).

Pour recommander des items de type textuel, le système établit son profil item qui est généralement décrit par un vecteur de mots clés. Comme dans le cas de *Fab system*

(Balabanovi & Shoham, 1997) qui recommande des pages Web, représentant le contenu d'une page Web par les plus importants 100 mots clés. L'importance d'un mot clé peut être déterminée par plusieurs mesures de pondération, et celles-ci peuvent être définies par différents moyens comme la mesure TF-IDF (*Term Frequency/Inverse Document Frequency*) (Salton, 1989).

## Force et faiblesses des méthodes traditionnelles :

La table ci-dessous résume les forces et faiblesses des méthodes traditionnelles, en l'occurrence le Filtrage Collaboratif (FC) et le Filtrage à Base de Contenu (CBF).

**Table 2.1** Forces et faiblesse des méthodes traditionnelles

| Technique | Forces | Faiblesses |
|---|---|---|
| CF | *Cross-genre niches* … (a)<br>Connaissance du domaine non requise … (b)<br>Adaptabilité : la qualité croit avec le temps … (c) | Problème du nouvel usager … (d)<br>Problème du nouvel item … (e)<br>Démarrage à froid (*Cold start*) … (f)<br>La *sparsity* … (g)<br>Le *gray Sheep* … (h)<br>Le *shilling* … (i) |
| CBF | (b)<br>(c) | (d)<br>Limite liée au contenu d'un item … (j)<br>*Overspecialization* … (k) |

Adapté de (Burke, 2002)

### (a) *Cross-genre niches*

Le filtrage collaboratif se distingue par sa capacité unique de recommander à un usager ce qui est hors du familier, c'est ce que Burke appelle : *cross-genre niches*. En effet, un usager peut se voir recommander des items de genres différents. Par exemple, un usager qui a des voisins similaires du point de vue des sports qu'ils favorisent et que ces voisins aiment la cuisine et les recettes. Même si cet usagé n'a jamais exprimé ce genre de favoris, il peut se voir recommander des recettes de cuisine.

***(b) Connaissance du domaine***

La connaissance du domaine n'est pas requise, le processus de recommandation se base uniquement sur les évaluations des items.

***(c) Adaptabilité***

Il y a une relation de proportion entre la qualité de la recommandation et la taille de la base de données des évaluations. En effet, Au fur à mesure que la base de données des évaluations augmente, la recommandation devient plus précise.

***(d) Problème du nouvel usager***

Un problème commun au filtrage collaboratif et au filtrage à base de contenu est qu'un nouvel utilisateur qui n'a pas encore accumulé suffisamment d'évaluations ne peut pas avoir de recommandations pertinentes. Cela est dû au fait que le système ne peut pas comprendre ses préférences d'où l'impossibilité de le classer ou de lui trouver un voisinage.

***(e) Problème du nouvel item***

C'est un problème qui concerne le filtrage collaboratif et non pas le filtrage à base de contenu. Car, dans le cas du filtrage à base de contenu, il suffit d'introduire l'item dans le système (la base de données) pour que celui-ci soit analysé et rentré dans le processus de recommandation. Alors que dans le cas du filtrage collaboratif il doit avoir suffisamment d'évaluations pour que celui-ci soit pris en considération dans le processus de recommandation. Contrairement au problème précèdent, le filtrage à base de contenu n'a pas besoin que l'item soit évalué par un grand nombre d'usagers comme dans le filtrage collaboratif.

***(f) Démarrage à froid (Cold start)***

On parle du problème du démarrage à froid quand on parle du filtrage collaboratif. Car ce problème concerne, aussi bien, le problème d'un nouvel usager, que le problème du nouvel item. On ne l'utilise pas dans le cas du filtrage à base de contenu, car le problème du nouvel item ne se pose pas (voir ci-dessus le paragraphe (e)). Il est difficile de faire des

recommandations pour un nouvel usager. Ou encore recommander un item nouvellement introduit. Le filtrage collaboratif exige un minimum d'évaluations pour fonctionner.

***(g) La Sparsity***

Ce problème survient quand il n'y a pas suffisamment évaluations d'items communs entre les usagers. Par exemple, le cas de quelques usagers qui évaluent des produits parmi un vaste choix et que ces groupes de produits ne se chevauchent pas. Ainsi, on ne peut pas établir des corrélations entre ces usagers.

(h) ***Le gray Sheep*** (Claypool et al., 1999)

Le filtrage collaboratif tend à recommander les items disposant des meilleures évaluations, ce qui crée un phénomène d'entrainement vers les items les plus populaires. En d'autres termes, plus un item est bien évalué, plus il est recommandé. C'est un peu comme : tout le monde aime ça, donc il y va de même pour lui.

***(i) Le Shilling***

C'est l'action malveillante d'influencer la recommandation en créant de faux profils pour voter et favoriser/défavoriser certains items. Par exemple, un vendeur qui, à travers des faux profils, valorise ses produits pour que le système les recommande pour les clients en ligne.

***(j) Limite liée au contenu d'un item***

La recommandation à base de contenu se base sur l'analyse des items. Cela n'est pas toujours évident. En effet, dépendamment de la nature ou du genre d'item, l'extraction de l'information est facile dans certains domaines comme la recherche d'informations appliquée à des documents, alors que cette approche est vague dans le cas des livres. Encore plus, si on considère le cas du contenu multimédia comme la musique, alors l'extraction de l'information est difficile.

***(k) Overspecialization***

C'est une limite au filtrage à base de contenu. Avec le filtrage à base de contenu, un usager ne peut pas avoir des recommandations autres que ce qu'il a connu déjà. En effet, la recommandation se base sur son profil, et tous les éléments hautement évalués dans ce dernier ne sont rien d'autre que ce qu'il avait connu avant. Une autre forme de ce problème est comment éviter de recommander des items très semblables à ce qu'il a déjà vu. En d'autres termes, comment juger qu'un item, semblable à ce que l'usager a déjà vu, puisse apporter suffisamment de nouveauté pour le lui proposer sans tomber dans la redondance ? En résumé, ce problème concerne la diversification du contenu, que ce soit en dépassant le genre (*Cross-genre niches point* (a)), soit en évitant la redondance.

Les problèmes ci-dessus ont fait l'objet de plusieurs études dans la littérature scientifique concernant les systèmes de recommandation. De nombreuses solutions sont proposées. Elles sont, généralement, des méthodes basées sur les deux méthodes traditionnelles présentées ci-dessus. Le paragraphe suivant présente un survol de ces méthodes dites *hybrides*.

## 2.3.3 Les méthodes hybrides

Par définition, un système hybride est une combinaison d'au moins deux techniques de recommandation pures (Burke, 2002) pour pallier les limites liées à chacune d'elles. Au moins deux méthodes pures, y compris les deux traditionnelles qui sont le filtrage collaboratif et le filtrage à base de contenu, car d'après ce dernier, il y a cinq méthodes pures. Alors que par rapport à (Adomavicius & Tuzhilin, 2005), l'approche hybride est la combinaison des deux méthodes traditionnelles précédentes. Ajouter à cela les méthodes basées sur d'autres techniques en provenance des statistiques et de l'apprentissage machine.

### 2.3.3.1 Les hybrides selon Burke

#### *2.3.3.1.1 Pondération (Weighted)*

Les résultats ou les votes qui sont générés par différentes techniques de recommandation sont combinés de façon à ce qu'une seule recommandation en résulte.

*2.3.3.1.2 Commutation (Switching)*

C'est une technique qui permet de faire le choix d'un modèle de recommandation parmi plusieurs, en se basant sur plusieurs critères. La détermination de la technique appropriée dépend de la situation. Le système se doit alors de définir les critères de commutation, ou les cas où l'utilisation d'une autre technique est recommandée. Ceci permet au système de connaître les points forts et les points faibles des techniques de recommandation qui le constituent.

*2.3.3.1.3 Technique mixte (Mixed)*

Cette technique est capable de donner à l'utilisateur des recommandations qui proviennent de plusieurs techniques (Smyth & Cotter, 2000). L'utilisation simultanée du filtrage collaboratif et du filtrage par contenu en est un exemple. Cette utilisation simultanée permet d'éviter les problèmes posés par le filtrage collaboratif, à savoir, le démarrage à froid. Aussi, le filtrage basé sur le contenu permet d'obtenir des recommandations sur de nouveaux objets, en se basant sur leurs descriptions respectives.

*2.3.3.1.4 Combinaison de caractéristiques (Features combination)*

Il s'agit d'une technique où les caractéristiques des informations qui sont fournies par les différentes méthodes de recommandation sont combinées, afin de permettre l'utilisation d'une technique unique sur l'ensemble des données. Le filtrage collaboratif utilise les votes des utilisateurs comme source de données. Ces données, lorsqu'ajoutées aux informations utilisées par le filtrage basé sur le contenu, peuvent produire des recommandations fiables.

*2.3.3.1.5 Cascade (Cascade)*

Cette méthode hybride se fait selon deux techniques :

- Une première technique permet de générer un ensemble de candidats potentiels.
- Une deuxième technique permet de raffiner les recommandations.

Cette méthode a pour avantage que la deuxième technique sert uniquement dans le cas de figure où les recommandations générées par la première nécessitent une discrimination additionnelle. Si la première technique génère peu de recommandations, ou si ces recommandations sont ordonnées afin de permettre une sélection rapide, la deuxième technique ne sera pas utilisée non plus.

*2.3.3.1.6 Augmentation de caractéristiques (Feature augmentation)*

Cette méthode ressemble à la méthode précédente (Cascade) du fait qu'une première technique est utilisée dans une première étape et les résultats produits sont utilisés dans une deuxième technique. Contrairement à la précédente, les données qui ont servi dans la première étape entrent avec ces résultats produits dans le processus de calcul de la deuxième étape.

*2.3.3.1.7 Métaniveau (Meta-level)*

Comme avec la précédente méthode (augmentation de caractéristiques), une première technique est utilisée, mais cette fois-ci, non pas pour produire de nouvelles caractéristiques, mais pour produire un modèle. Et dans la deuxième étape, c'est tout le modèle qui servira d'entrée pour la deuxième technique.

### 2.3.3.2 Les Hybrides selon Adomavicius et Tuzhilin

(Adomavicius & Tuzhilin, 2005) distinguent quatre façons de combiner les deux méthodes précédentes à savoir :

- Implémenter la méthode collaborative et la méthode basée sur le contenu séparément puis combiner leurs prédictions.
- Incorporer quelques caractéristiques de la méthode basée sur le contenu dans l'approche collaborative.
- Incorporer quelques caractéristiques de la méthode collaborative dans l'approche à base de contenu.

- Construction d'un modèle général unifié qui incorpore les caractéristiques des deux modèles.

*2.3.3.2.1 Combinaison séparée de méthodes de recommandation.*

Cette approche hybride consiste en une implémentation séparée des deux méthodes de base. Par la suite, les résultats obtenus séparément par chacune de ces méthodes seront combinés linéairement pour constituer un seul résultat comme dans les articles suivants (Claypool et al., 1999; Pazzani, 1999). Une autre façon de combiner les deux méthodes est de choisir le meilleur résultat en se basant sur certaines mesures de qualité. Par exemple, (Tran & Cohen, 2000) qui choisit celui dont la recommandation est plus compatible avec les estimations passées de l'utilisateur.

*2.3.3.2.2 Ajout de caractéristiques de la méthode basée sur le contenu dans l'approche collaborative.*

La similarité entre deux usagers est calculée en se basant sur leur profil basé sur le contenu. Différemment du filtrage collaboratif pur qui utilise directement les évaluations des usagers pour trouver la similarité entre deux usagers, cette méthode hybride utilise ces mêmes évaluations combinées avec le contenu caractéristique des items pour trouver le profil de chaque usager. Une fois que les profils basés sur le contenu sont déterminés, la méthode de filtrage collaboratif est appliquée sur ces derniers pour recommander les items.

Plusieurs applications de cette méthode dans la littérature scientifique notamment (Pazzani, 1999) qui l'ont dénommé « *Collaborative via content* » que nous traduisons par « collaboration à travers le contenu », on la retrouve aussi dans le système Fab de (Balabanovi & Shoham, 1997). L'avantage de cette méthode, comme mentionné par (Pazzani, 1999), est de contourner les problèmes liés aux espacements.

*2.3.3.2.3 Ajout de caractéristiques collaboratives dans l'approche à base de contenu.*

La plus populaire approche de cette catégorie est l'utilisation de quelques techniques de réduction de la dimensionnalité dans un groupe de profils basés sur le contenu. Par exemple (Soboroff & Nicholas, 1999) utilise LSI (*latent semantic indexing*) pour créer une vue collaborative d'une collection de profils d'usager, où les profils d'usagers sont représenté par des vecteurs limite (comme discuté dans la section 2.1), le résultat est l'amélioration de la performance comparativement à l'approche purement basée sur le contenu.

*2.3.3.2.4 Développement d'un modèle de recommandation unifié et unique.*

Cette catégorie regroupe toutes les méthodes qui combinent en un seul et unique modèle les deux traditionnelles méthodes pures qui sont le filtrage collaboratif et le filtrage à base de contenu. En effet un même algorithme ou une même formule regroupent les deux méthodes comme l'algorithme (Basu, Hirsh, Cohen, & Manning, 2001). Plusieurs modèles ont été implémentés, par exemple (Basu, Hirsh, Cohen, & Manning, 2001) propose une unification dans un classificateur à base de règles. De même, (Popescul, Ungar, Pennock, & Lawrence, 2001; Schein, Popescul, Ungar, & Pennock, 2002) proposent un modèle probabilistique unifié pour combiner les deux méthodes, lequel est basé sur l'analyse sémantique latente probabilistique (*probabilistic latent semantic analysis*).

### 2.3.3.3 Le filtrage à base de connaissances

Cette classe de méthodes incorpore des techniques de l'apprentissage machine, plus précisément, toutes les techniques d'acquisition de la connaissance du domaine de l'intelligence artificielle. Comme le raisonnement à base CBR (*Case-Based Reasoning*) utilisé par Burke dans son système de recommandation de restaurants : Entrée (Burke, 2000). Ce système exploite une bonne connaissance du domaine, d'ailleurs Burke dans sa classification lui a dédié, une classe à part entière qu'il appelle : les systèmes de recommandation basés sur la connaissance (en anglais *Knowledge-based recommender system*). Généralement, ces types de systèmes sont utilisés là où le domaine est disponible

sous une certaine structure facilement compréhensible par la machine comme dans le cas des systèmes *Quickstep* et *Foxtrot* (Middleton, Shadbolt, & De Roure, 2004) qui recommandent des articles de recherche en se basant sur une *Ontologie* de sujet d'articles de recherche.

**Résumé de la classification**

Les systèmes de recommandation se classent en trois catégories admises par la majorité des chercheurs de ce domaine : le filtrage collaboratif, le filtrage à base de contenu et les méthodes hybrides. Burke a ajouté trois autres : le *filtrage démographique*, le *filtrage à base d'utilité* et le *filtrage à base de connaissance*. Le premier est un cas particulier du filtrage collaboratif. Quant aux deux autres, le filtrage à base d'utilité est un cas particulier du filtrage à base de connaissance. Pour cette dernière, (Adomavicius & Tuzhilin, 2005) la considère comme une méthode Hybride. Nous la considérons de même, pour deux raisons : le fait de la distinguer comme méthode pure brise le consensus. La deuxième raison est : l'approche de (Adomavicius & Tuzhilin, 2005) est plus générale et bien formulée. De ce fait, la classification de Burke se ramène à cette dernière. Concernant les sous-classes des Hybrides, ils s'entendent sur le fait qu'elles sont un mélange entre les deux classiques et pures, alors que (Adomavicius & Tuzhilin, 2005) les ont étendues pour leur associer toute autre méthode, même si elle n'est pas composée de celles dites pures. D'ailleurs, la méthode de filtrage à base de connaissance est considérée comme telle. La logique de classification proposée par (Adomavicius & Tuzhilin, 2005) voir Table 2.2 est meilleure par rapport à celle de (Burke, 2002) pour les raisons suivantes :

- Formulation générale du problème de recommandation (Paragraphe 2.2.2)
- Séparation entre l'implémentation (algorithme) et les techniques de recommandation (Table 2.2).
- Mise en évidence de la symétrie entre la recommandation usager à usager et item à item. (Paragraphe 2.2.2).

Néanmoins, la terminologie, introduite par (Burke, 2002) surtout en ce qui concerne les méthodes hybrides, demeure prédominante; chose attribuable à son ancienneté et à l'intuitivité du nommage utilisé.

**Table 2.2** Classification selon (Adomavicius & Tuzhilin, 2005).

| Recommendation Approach | Recommendation Technique | |
|---|---|---|
| | Heuristic-based | Model-based |
| Content-based | Commonly used techniques:<br>• TF-IDF (information retrieval)<br>• Clustering<br>Representative research examples:<br>• Lang 1995<br>• Balabanovic & Shoham 1997<br>• Pazzani & Billsus 1997 | Commonly used techniques:<br>• Bayesian classifiers<br>• Clustering<br>• Decision trees<br>• Artificial neural networks<br>Representative research examples:<br>• Pazzani & Billsus 1997<br>• Mooney et al. 1998<br>• Mooney & Roy 1999<br>• Billsus & Pazzani 1999, 2000<br>• Zhang et al. 2002 |
| Collaborative | Commonly used techniques:<br>• Nearest neighbor (cosine, correlation)<br>• Clustering<br>• Graph theory<br>Representative research examples:<br>• Resnick et al. 1994<br>• Hill et al. 1995<br>• Shardanand & Maes 1995<br>• Breese et al. 1998<br>• Nakamura & Abe 1998<br>• Aggarwal et al. 1999<br>• Delgado & Ishii 1999<br>• Pennock & Horwitz 1999<br>• Sarwar et al. 2001 | Commonly used techniques:<br>• Bayesian networks<br>• Clustering<br>• Artificial neural networks<br>• Linear regression<br>• Probablistic models<br>Representative research examples:<br>• Billsus & Pazzani 1998<br>• Breese et al. 1998<br>• Ungar & Foster 1998<br>• Chien & George 1999<br>• Getoor & Sahami 1999<br>• Pennock & Horwitz 1999<br>• Goldberg et al. 2001<br>• Kumar et al. 2001<br>• Pavlov & Pennock 2002<br>• Shani et al. 2002<br>• Yu et al. 2002, 2004<br>• Hofmann 2003, 2004<br>• Marlin 2003<br>• Si & Jin 2003 |
| Hybrid | Combining content-based and collaborative components using:<br>• Linear combination of predicted ratings<br>• Various voting schemes<br>• Incorporating one component as a part of the heuristic for the other<br>Representative research examples:<br>• Balabanovic & Shoham 1997<br>• Claypool et al. 1999<br>• Good et al. 1999<br>• Pazzani 1999<br>• Billsus & Pazzani 2000<br>• Tran & Cohen 2000<br>• Melville et al. 2002 | Combining content-based and collaborative components by:<br>• Incorporating one component as a part of the model for the other<br>• Building one unifying model<br>Representative research examples:<br>• Basu et al. 1998<br>• Condliff et al. 1999<br>• Soboroff & Nicholas 1999<br>• Ansari et al. 2000<br>• Popescul et al. 2001<br>• Schein et al. 2002 |

Nous nous baserons principalement dans le reste de ce mémoire sur la classification de (Adomavicius & Tuzhilin, 2005) et nous ferons référence à celle de (Burke, 2002) pour une meilleure clarté. Dans ce qui suit, nous poursuivons en présentant deux extensions proposées par (Adomavicius & Tuzhilin, 2005) pour les systèmes de recommandation à

savoir les notions de *multi dimensionnalité* et de *multicritère* de la recommandation. Nous prêterons plus d'attention au multicritère, car c'est l'une des cibles de notre contribution.

## 2.4 Les systèmes de recommandation multidimensionnels

Les systèmes de recommandation discutés ci-haut opèrent dans un espace bidimensionnel *Usager x Item* du fait qu'ils tiennent compte que des deux profils : usager et item. Alors que d'autres informations *contextuelles* qui peuvent s'avérer pertinentes ne sont pas prises en considération. Cependant, dans plusieurs situations, l'utilité de certains produits pour un usager dépend étroitement du temps (par exemple la saison, le soir, la nuit) et par fois elle dépend des personnes avec qui le produit sera consommé et sous quelles circonstances. La recommandation traditionnelle d'items pour des usagers, dans ce cas, n'est pas suffisante, elle doit prendre en considération les informations contextuelles. Par exemple, dans la recommandation de films, on doit tenir compte des personnes avec qui l'usager projette de voir le film. Car le cas diffère selon qu'il voudra le voir avec ses amis, ses parents ou ses enfants. Le temps aussi peut avoir un effet, comme dans le temps des fêtes il pourra préférer des films en rapport avec cette période. Un autre exemple, la recommandation d'habits est fortement dépendante de la saison, de l'objectif (costume de fête, de sport, de déguisement, etc.).

Sur le plan application, plusieurs des méthodes bidimensionnelles classiques ne s'appliquent pas directement dans un cas multidimensionnel (Adomavicius, Sankaranarayanan, Sen, & Tuzhilin, 2005; Adomavicius & Tuzhilin, 2001). En effet, il y a un besoin de les adapter. Par exemple, (Adomavicius, Sankaranarayanan, Sen, & Tuzhilin, 2005) propose la recommandation basée sur la réduction de la dimension de l'espace de recommandation (*reduction-based recommendation aproach*). Cette approche filtre le profil d'un usager en tenant compte du contexte courant. Par conséquent, le système utilise uniquement les évaluations des usagers relatives à un contexte précis.

## 2.5 Les Systèmes de Recommandation Multicritère

La plupart des systèmes de recommandation courants opèrent sur des évaluations monocritères. Un critère général et unique qui exprime si l'usager aime ou n'aime pas

l'item. Il est indiqué par une valeur unique choisie parmi une échelle de valeurs données. Alors que, deux usagers qui évaluent un item de la même façon, cela n'implique pas qu'ils l'ont fait pour les mêmes raisons. En effet, si l'on considère un exemple d'évaluation de restaurants, deux usagers peuvent donner la même note : « Très bon » pour un tel restaurant, alors que : un des deux l'aime pour sa spécialité en cuisine, alors que l'autre l'aime pour son ambiance et sa localisation au bord de la mer. Pour bien illustrer cette idée, nous allons reprendre un exemple détaillé extrait d'un article de journal (Adomavicius & Kwon, 2007).

**Exemple détaillé**

Considérons un système de recommandation de films où les usagers évaluent des films suivant une échelle allant de 1 jusqu'à 13. La valeur : 1 représente la note la plus mauvaise et 13 la meilleure note. Rappelons que le filtrage collaboratif revient à trouver les usagers les plus similaires à un usager cible (son voisinage) afin de lui recommander les items hautement appréciés par ses derniers et qu'il n'a pas encore évalués.

Considérons les deux cas : cas d'un système de recommandation monocritère basé sur le filtrage collaboratif et le deuxième cas, un système de recommandation multicritère basé aussi sur le filtrage collaboratif.

**Cas de la recommandation monocritère (traditionnelle) :**

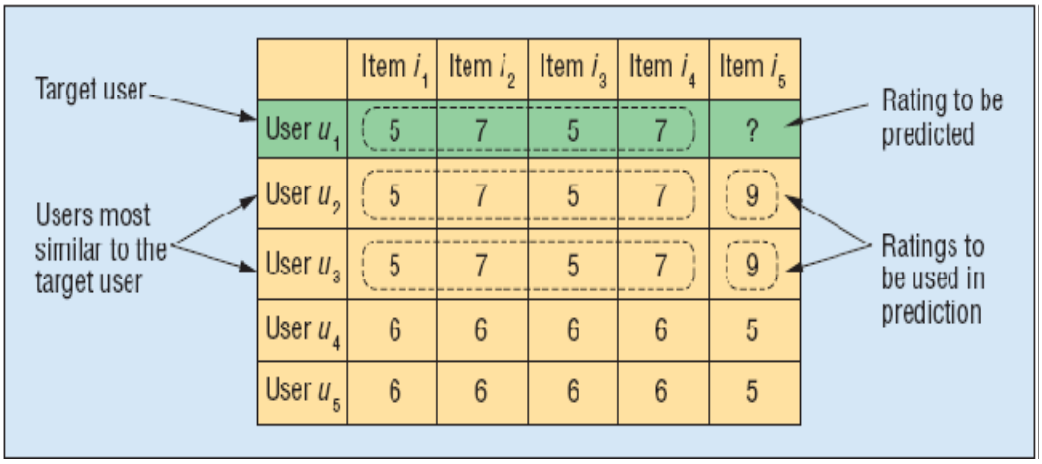



**Figure 2.1** Exemple d'évaluation monocritère (Adomavicius & Kwon, 2007)

La Figure 2.1 illustre ce cas. Soit l'usager *User* $U_1$ l'usager cible à qui la recommandation est destinée. Nous voyons bien que les deux usagers qui lui sont plus similaires en terme d'évaluation sont *User* $U_2$ et *User* $U_3$. L'item $I_5$ étant hautement apprécié par ces deux voisins, donc le système va conclure qu'il sera aussi hautement apprécié par l'usager cible d'où la recommandation de cet item.

**Cas de la recommandation multicritère :**

Cette fois ci, les usagers $U_1 \ldots U_2$ évaluent les films $i_1 \ldots i_5$ suivant quatre critères différents, de la même manière que le site Web YAHOO! MOVIES (URL, 14). Les usagers évaluent les films suivant quatre critères : histoire, action, direction et visuels. La Figure 2.2 illustre ce cas.

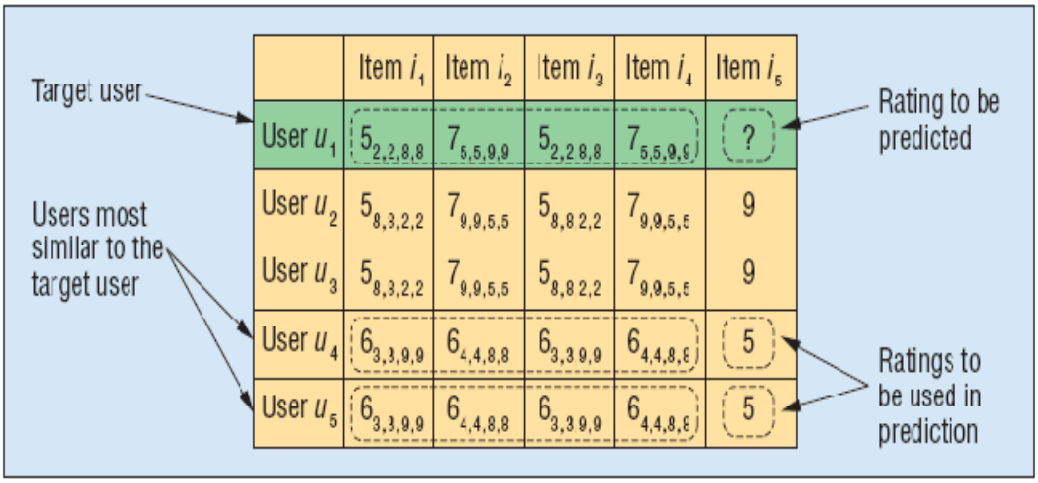



**Figure 2.2** Exemple d'évaluation multicritère(Adomavicius & Kwon, 2007)

Nous voyons bien, cette fois-ci, que les deux usagers semblables à l'usager cible User $U_1$ sont $U_4$ et $U_5$, malgré leur différence dans l'évaluation générale. En effet, les différents critères d'évaluation nous ont révélé des détails cachés par rapport au premier cas, ce qui nous a permis de comprendre le pourquoi des évaluations monocritères. Ainsi, le système prédira que l'usager $U_1$ donnera une valeur 5 pour l'item $i_5$. Pour récapituler, l'évaluation générale montre à quel point un usager aime tel produit. Alors qu'une évaluation détaillée montre le pourquoi derrière l'appréciation du produit. Ainsi, cet exemple démontre la force et la pertinence d'un système de recommandation multicritère, comparativement à un système traditionnel. À ce stade, pour prendre en charge l'aspect

multicritère, (Adomavicius & Kwon, 2007) ont proposé des approches pour étendre les algorithmes traditionnellement monocritères. Nous considérons dans ce mémoire, l'approche visant l'extension du filtrage collaboratif, intitulé : « *Aggregating traditional similarities from individual criteria* ». Cette approche discute des façons de combiner les similarités individuelles qui résultent du filtrage collaboratif traditionnel pour avoir une seule similarité. Cette dernière sera utilisée dans la formule de prédiction traditionnelle.

**Extension du filtrage collaboratif traditionnel pour le cas multicritère**

Les auteurs de cette approche ont proposé à titre d'exemple deux façons, parmi d'autres, de combiner les similarités individuelles en un seul vecteur pour calculer le voisinage. La première est une simple moyenne et la deuxième calcule le minimum des similarités.

Soit : $U = \{u_1, u_2, \dots, u_k\}$ ensemble des $k$ usagers

$S = \{s_1, s_2, \dots, s_l\}$ ensemble des $l$ items

Chaque item est évalué suivant $n$ critères $c_i$

Soit :

$C = \{c_1, c_2, \dots, c_n\}$ l'ensemble de ces $n$ critères.

Le calcul de la similarité multicritère passe par le calcul des similarités individuelles exactement comme le cas de la similarité traditionnelle dite monocritère. C'est-à-dire, nous calculons la similarité entre deux usagers $u$ et $u'$ notée $sim_i(u, u')$ par rapport à chaque critère $c_i$ considéré seul avec n'importe quelle formule s'appliquant dans le cas traditionnel. Nous obtenons :

- La *moyenne de similarités* par la formule suivante :

$$sim_{avg}(u, u') = \frac{1}{k+1} \sum_{i=1}^{k} sim_i(u, u')$$

Avec : $sim_i$ = *{ensemble des similarités par rapport au critère i}*

- Le *minimum des similarités* est donné par la formule suivante:

$$sim_{min}(u, u') = \min_{i=1,\dots k} sim_i(u, u')$$

Cette dernière approche considérer le minimum des similarités pour déterminer le voisinage de prédiction.

Dans ce mémoire, nous allons étudier plusieurs façons de combiner les similarités pour trouver le voisinage optimal à utiliser pour prédire les évaluations pour chacun des critères.

## 2.6 Conclusion

Après avoir présenté les deux approches principales dans la classification des systèmes de recommandation, nous avons pris position en adoptant la classification de (Adomavicius & Tuzhilin, 2005). Rappelons les raisons principales de ce choix :

- Formulation générale du problème de la recommandation
- Séparation entre l'implémentation (algorithme) et les techniques de recommandation (Table 2.2).
- Mise en évidence de la symétrie entre la recommandation Usager à Usager et Item à Item.

Sans pour autant négliger l'autre approche (Burke, 2002), nous ferons appel à sa terminologie concernant les méthodes hybrides, car elle bénéficie d'une large utilisation dans ce domaine et pour une meilleure clarté. Nous avons présenté les notions de multidimensionnalité et de multicritère des systèmes de recommandation que les auteurs (Adomavicius & Tuzhilin, 2005) ont proposés parmi les extensions possibles de ces systèmes. Nous ajoutant à cette occasion une autre raison pour avoir choisi la classification précédente, et c'est une meilleure conceptualisation pour supporter les nouvelles tendances entre autres la multidimensionnalité et le multicritère. Nous avons prêté une attention particulière pour la notion de multicritère dont nous avons présenté des travaux issus de (Adomavicius & Kwon, 2007), car cela nous servira comme base pour présenter nos

algorithmes (Naak, Hage, & Aïmeur, 2009) testés dans le cadre de ce mémoire pour améliorer la précision de la recommandation.

Le prochain chapitre présente le contexte de notre recherche ainsi que les problématiques que nous proposons de résoudre. Particulièrement, la dernière partie de ce chapitre propose un état de l'art des *systèmes de recommandation d'articles de recherche*. À notre connaissance, nos approches multicritère n'ont jamais été explorées dans ce domaine d'application, et de ce fait elles constituent une des contributions de ce mémoire.

# Chapitre 3 Les systèmes de gestion d'articles de recherche

Ce chapitre approfondit l'objet central de notre recherche, qui est l'article avec ses méta-informations, dans l'optique de bien comprendre ses spécificités et ainsi mieux cerner les exigences de la recherche. Après une analyse des besoins liés à ces articles, nous présentons des solutions existantes pour tenter de les satisfaire. Plus précisément, nous passons en revue trois classes de systèmes : les systèmes de gestion bibliographique, les systèmes de gestion de contenu et finalement les systèmes de recommandation d'articles de recherche. Nous constatons que ces solutions ne sont pas adaptées pour les besoins spécifiques de ce qui est l'objet de notre recherche.

## 3.1 L'article de recherche

L'article de recherche scientifique est souvent un couronnement de travaux de recherches d'un ou plusieurs chercheurs, à travers lequel, ils publient leurs résultats qui ne sont pas nécessairement, des créations ou des découvertes, mais aussi des synthèses de revue de la littérature scientifique d'un domaine bien défini. En effet, rassembler et analyser exhaustivement plusieurs articles de recherche relevant d'un certain sujet et apporter des critiques et de nouvelles orientations est une grande tâche. Ces derniers types d'articles sont appelés *Survol de littérature* (*Survey*). Leur apport dans le domaine de la recherche est d'une très grande importance, du fait de leur nature qui est de comprendre un état de l'art des plus complets et profonds d'un domaine de recherche. Par ailleurs, parmi les plus importantes caractéristiques de l'article de recherche, Nous citons son *originalité* et sa *contribution* à son domaine de recherche. D'après le grand dictionnaire terminologique[1], l'originalité est la « Qualité d'une œuvre de l'esprit qui paraît être réellement différente, tant

[1] Grand dictionnaire terminologique, Office québécois de la langue française : http://www.granddictionnaire.com/btml/fra/r_motclef/index800_1.asp

par son contenu que par sa forme, de toute œuvre analogue » et « L'originalité peut exister sans qu'elle ait trait à quelque chose de nouveau ». La *contribution*, c'est l'apport concret dans un domaine comme l'amélioration d'une méthode, la résolution d'un problème. La communauté scientifique s'attend qu'un article ait au moins une contribution convaincante à son domaine de recherche.

Le processus pour la reconnaissance de cette recherche par la communauté scientifique passe par plusieurs étapes et l'article doit satisfaire plusieurs conditions avant sa publication dans une revue scientifique ou un procédé. La valeur de l'article, mis à part son contenu, est influencée par la crédibilité de la conférence, mais aussi par plusieurs critères de contenu et de forme. Le contenu comprend entre autres son originalité et sa contribution, sa pertinence pour la conférence en question, sa qualité technique. Pour les critères de forme, l'article doit répondre aux exigences de format de la conférence, tels qu'une mise en page en double colonne, suivre le style bibliographique proposé, etc.

### 3.1.1 Structure de l'article de recherche

En général, un article de recherche est composé d'un *titre*, l'indication de l'*auteur* ou des *auteurs*, un *résumé* (*abstract*), les *mots clés*, une *introduction*, un *état de l'art*, une *présentation de la solution*, une *conclusion* et une *liste de références* ou une *bibliographie*. Nous décrivons brièvement chacun de ces éléments. Le *titre* requiert une attention particulière par rapport au reste des sous-titres. En effet, il doit être bien choisi de sorte à refléter un des aspects de cette recherche. La mention de l'*auteur* ou des *auteurs* de cette recherche est accompagnée, en général, de leurs coordonnées : adresses et courriels. Le résumé donne une vue générale de l'article. L'introduction présente la problématique ainsi que le contexte ou le domaine de la recherche. L'état de l'art résume et discute les travaux similaires pour montrer l'originalité des travaux en question. La présentation de leur solution peut contenir une partie introduisant les prés requis permettant de comprendre la solution développée, il peut s'agir de notions élémentaires ou de travaux préalables sur lesquels est basée la présente recherche. La présentation des résultats ainsi qu'une discussion de ces derniers. Une conclusion, pour récapituler le tout. Et finalement, tout article doit inclure une bibliographie pour toutes les références citées par les auteurs de

l'article en question. La présence d'un résumé, d'une discussion détaillée des résultats et d'une bibliographie conséquente sont de bons indicateurs du caractère scientifique d'un article.

### 3.1.2 Les types d'articles

Dans un domaine de recherche, un article peut être classé en suivant plusieurs critères. Néanmoins, Nous nous intéressons dans ce paragraphe à la classification suivant l'objectif de l'article. Notons qu'il n'existe pas de consensus pour un tel classement, mais en général, la majorité des conférences par arbitrage (*peer reviews*) considère trois catégories non exhaustives : *Recherche*, *Rapport technique* et *survol*. L'objectif d'un article de type recherche est d'analyser et de proposer des solutions rationnelles aux problèmes à la fine pointe de la recherche dans le domaine. Il peut être théorique, sans s'intéresser à l'aspect expérimental et de son application concrète, ou expérimental, ce qui induit un aspect pratique et des données expérimentales pour valider les résultats et les conclusions trouvées. Un article de type rapport est en général une étude de cas qui peut prendre plusieurs formes, par exemple : une étude et une critique de l'existant dans le but de proposer des améliorations, ou une étude comparative pour prouver l'apport et efficacité d'une solution proposée. Quant à l'article de type survol, c'est l'étude et l'analyse d'une vaste liste bibliographique pour dresser le constat de ce qui est l'état de l'art d'un domaine. En d'autres termes, c'est une synthèse de revues de littérature scientifique d'un domaine bien défini. Les auteurs d'un tel type d'articles rassemblent, et analysent exhaustivement plusieurs articles de recherche relevant d'un certain sujet et apportent des critiques et de nouvelles orientations. Leur apport dans le domaine de la recherche est d'une très grande importance, par le fait de leur nature de comprendre un état de l'art des plus complets et profonds d'un domaine de recherche.

### 3.1.3 La Référence, la Citation et le Style de bibliographie

Les articles de recherche, une fois publiés, sont protégés par des droits d'auteur. Cela concerne, entre autres, le texte et les idées véhiculées. À défaut de mention de la source d'un extrait ou la reprise d'une idée sans citer son origine, l'auteur en question

risque d'être accusé de *plagiat*. La citation des ressources utilisées est une reconnaissance de la propriété intellectuelle et elle a de nombreux avantages pour la recherche en cours, par exemple : s'en servir pour argumenter, appuyer, renforcer et prouver les différentes idées proposées. Elles serviront aussi de renvoi vers d'autres articles et, par conséquent, un lecteur voulant plus de détails ou en quête de recherche similaire se verra rassasié.

Il est important de fixer certains termes pour éviter des confusions à commencer par l'article de recherche lui même. Quand nous parlons de celui-ci, cela peut induire deux concepts qui sont en réalité inhérents : le *texte de l'article* et son *identifiant*. Le premier désigne le contenu de l'article, nous nous y référons par le terme *copie physique* ou *copie digitale*, il pourra exister sous divers formats, le plus répondu est le format PDF (*Portable Document Format*). Pour le deuxième, un article est désigné par sa citation ou référence. C'est ce que font remarquer les auteurs dans l'article (Torres, McNee, Abel, Konstan, & Riedl, 2004) "*We will draw a subtle but important difference between a paper and a citation. A citation is a paper for which the text may not be available. A citation therefore is a pointer to a paper. On the other hand, a paper is a citation for which we also have its text. This is important because many citations may be references to papers that we do not have in digital format.* "

Dans un contexte de publication, la citation d'une ressource académique est concrétisée par trois notions : le *pointeur de la citation* ou tout simplement *citation*[1], la *référence* et troisièmement la *bibliographie*. La *citation* est un pointeur, vers la *référence* dans la *bibliographie* située à la fin du document. La *référence* est une suite de métadonnées identifiant les articles cités. L'ensemble de ces références regroupé à la fin du document constitue la *bibliographie*. Quant au style bibliographique, celui-ci représente le comment de l'écriture de ces trois informations. On parle de comment formater :

- la citation ? : il existe plusieurs formats pour la citation, mais il y a deux qui sont les plus répondus, la numérotation et le couple (auteur, année). La numérotation, par

[1] Dans un contexte de gestion de références, ce terme « *citation* » signifie « *pointeur de la référence* » ou encore « *pointeur de la citation* ». C'est une note qu'un auteur met devant le texte à citer et qui pointe vers la référence située à la fin du document. Par contre dans la littérature scientifique, les deux termes *citation* et *référence* sont employés indifféremment pour désigner l'un ou l'autre.

exemple : [1] [2] et [1,2] sont deux écritures différentes pour deux citations différentes. Le couple (auteur, année de publication), où plusieurs écritures sont possibles;

- la référence ? : il s'agit de combiner trois critères d'apparition pour les métadonnées à inclure dans la référence. Le premier est lesquelles de ces métadonnées il faut inclure ? Dans quel ordre ? Et quelles ponctuations utiliser pour les séparer ?
- La bibliographie ? : à la fin du document on retrouve la liste ordonnée de toutes les références qui constituent la bibliographie du document. Il faut choisir l'ordre de cette liste (ordre alphabétique, ordre d'apparition dans le texte, etc.)

## 3.2 Analyse des besoins dans le domaine de la recherche

Comme ce qui est dit précédemment, l'article est l'aboutissement d'une longue recherche et un couronnent d'un énorme effort. Cependant, pour y arriver les chercheurs sont souvent en perpétuelle prospection de nouveautés dans leur domaine. Cette prospection est faite principalement à travers la lecture de plusieurs et différents articles de recherches. De nos jours, la publication et la diffusion d'articles, comme toute autre information, sont rendues faciles et rapides grâce aux Nouvelles Technologies de l'Information et de la Communication (NTIC), principalement par le biais de l'Internet. En effet, la production d'articles sur support numérique à permis un très grand rendement dans la production et une haute accessibilité notamment avec l'avènement de gigantesques *bibliothèques numériques*, nous citons à titre d'exemple : Book24x7[1] pour les livres électroniques, IEEE (URL, 1) et ACM (URL, 2) pour les bases de données d'articles de recherche. Basées sur le Web, elles renferment des milliers d'articles sous formats numériques. Malgré les nombreux avantages dus à cette évolution, d'autres problèmes sont engendrés, notamment le fait que les chercheurs se trouvent submergés dans une marée d'informations qui ne cesse de s'amplifier.

[1] http://www.books24x7.com

Certes, les bibliothèques numériques nous offrent un accès instantané aux ressources électroniques, mais la gestion de ces dernières et leurs localisations sont devenues des tâches plus compliquées. Cela a engendré un effet inverse qui réduit la productivité sinon la qualité de cette production. Ce phénomène est dû à la perte et l'ignorance d'informations précieuses comme l'omission de mentionner un travail pertinent pour une recherche, à cause tout simplement, d'une recherche incomplète dans l'étape de la construction de l'état de l'art. Ces problèmes se traduisent par un besoin d'applications pour réduire la charge supportée par ces usagers. Nous avons recensé trois catégories de besoins : les *besoins de gestion bibliographique*, *besoin de gestion de la connaissance* et la *localisation de ressources* à savoir les articles de recherche ainsi que les connaissances qui leur sont liées. Nous procéderons par l'analyse de ces besoins avant d'étudier les solutions existantes.

### 3.2.1 Les besoins de gestion bibliographique

La production d'articles de recherche oblige la citation des autres travaux pour différentes raisons, dont certaines ont été énoncées au paragraphe 3.3.1, et parfois les travaux, même de l'auteur en question pour par exemple dire que les travaux courants sont basés sur une publication précédente. Bref, tout au long de leur carrière, les chercheurs auront souvent à manipuler ces références et ces citations. Pour des fins de publication, ces références doivent respecter un certain style exigé par le publicateur. Par exemple, si nous prenons un article qui sera publié par ACM, le style à respecter pour écrire une référence est comme suit :

[1] Naak, A., Hage, H. and Aïmeur, E. 2008. Papyres: a Research Paper Management System. *In Proceedings of the IEEE Joint Conference on E-Commerce Technology (CEC 08) and Enterprise Computing, E-Commerce and E-Services (EEE 08)*, (Crystal City, Washigton, D.C., USA, 21- 24 July, 2008).

**Figure 3.1** Exemple d'une référence formatée

- Pour la citation d'un article dans le texte, on utilise un numéro mis entre deux crochets et ces numéros suivent l'ordre alphabétique de la bibliographie. Dans l'exemple ci-dessus la citation est : [1].
- La référence est constituée respectivement par les éléments suivants : la liste des auteurs, le titre de l'article, nom de la conférence, l'endroit où cette conférence s'est déroulée, et finalement la date de cet événement.
- Le format doit être respecté : la façon d'abréger le prénom de chacun des auteurs, la ponctuation et l'endroit où elle est placée, l'ordre des éléments, etc.
- Pour la liste des références, elle doit respecter l'ordre alphabétique.

La manipulation de ces références, comme l'introduction ou la suppression d'une référence, perturbe cette organisation, ce qui rend cette tâche fastidieuse. Cette tâche est souvent répétée tout au long de la carrière du chercheur, même si la citation et souvent répétée dans différents contextes, la tâche demeure différente, car le style de référence dépend du contexte. En effet, avec tous les divers formats de citation, il devient difficile et ennuyant de formater et d'adapter les références à chaque fois que le format change. Décharger les chercheurs de ces tâches routinières les laissera libres de se concentrer sur le fond de leurs travaux, ce qui augmentera la productivité et la qualité intellectuelle. En effet, l'utilisation d'un système qui facilite la gestion de références comme des outils pour formater automatiquement les références selon différents styles, ainsi que les différentes fonctionnalités comme importer et exporter des références rendront la tâche facile pour les chercheurs. Plusieurs solutions sont proposées pour gérer les références, nous allons les étudier dans les prochains paragraphes et nous donnerons des études de cas d'application.

### 3.2.2 Les besoins de gestion de documents

L'article de recherche en format numérique est représenté par un fichier sauvegardé et classé dans un répertoire ou sous-répertoire se trouvant sur un support d'enregistrement tel qu'un disque dur. L'organisation et la gestion de ces fichiers dans un bas niveau en utilisant le gestionnaire de fichiers, occasionne plusieurs problèmes. Par exemple, l'utilisation de répertoire et sous-répertoire, pour classer les articles suivant plusieurs

critères, engendre des copies redondantes et une difficulté de gestion : la suppression, la modification du nom de fichier pour un article donné impose de parcourir tous les répertoires et sous-répertoires. Que dire de vouloir associer les différentes métadonnées (informations de citation, commentaires, résumés, etc.) avec le texte de l'article et vouloir faire la classification et le tri suivant les différents critères. Ces problèmes frôlent l'impossible à grande échelle, lorsque la liste de ressource à gérer est très longue et la structure de répertoires et sous-répertoires est complexe. C'est pourquoi, le besoin d'un système de gestion de documents et indispensable. Ainsi, ce dernier facilite l'organisation et la gestion et supporte d'autres fonctionnalités très avancées comme le suivi de l'état d'un article qui est d'une grande aide pour les chercheurs, ils le renseignent sur des aspects subjectifs entourant ses articles, comme : lu/non lu, commenté/non commenté, revu/non revu.

Le partage et la collaboration sont aussi d'une grande importance dans le domaine de la recherche. Il est souhaitable d'avoir une plate-forme d'échange de documents et de connaissances entre les membres participants. De façon plus générale, l'introduction des aspects Web 2.0 (O'Reilly, 2005) est d'une grande utilité notamment les étiquettes (*Tags*) les forums de discussions, le courriel, les flux RSS,[1]. Tout cela augmentera la productivité et la connaissance des chercheurs. Nous verrons dans le paragraphe concernant les systèmes de gestion de documents comment cet aspect est développé.

### 3.2.3 Les besoins de localisation de ressources

La localisation de ressources dans les gigantesques banques de données devient de plus en plus difficile, surtout avec la grande vitesse à laquelle celles-ci augmentent. La rapidité avec laquelle un chercheur retrouve ces articles est un facteur important pour son efficacité et son rendement. Nous distinguons deux besoins différents dans la localisation de ressources scientifiques. Le premier est la recherche basée sur les critères spécifiés explicitement par le chercheur, et le deuxième, est une localisation basée sur un système

[1] RSS, un acronyme qui a évolué à travers des versions : RSS 0.91 (*Rich Site Summary*), RSS 0.90 et 1.0 (*RDF Site Summary*) et RSS 2.0 (*Really Simple Syndication*).

intelligent capable de comprendre certains goûts du chercheur et lui recommander les ressources adéquates; Nous parlons, dans ce cas, de *localisation personnalisée* ou *système de recommandation*.

La localisation dans le premier cas est indépendante de l'usager. En d'autres termes, quel que soit l'usager, s'il utilise les mêmes critères le résultat est toujours le même. La complexité d'un tel système est variable, car il peut s'agir d'une simple recherche basée sur des critères bien définis d'un article. En général, les informations liées à un article sont gardées dans des bases de données bien structurées, et la recherche par rapport à l'un de ces critères revient, tout simplement, à envoyer des requêtes vers le Système de Gestion de Bases de Données (SGBD). Par exemple, les chercheurs ont souvent besoin de localiser dans leurs propres ressources les articles écrits par un certain auteur, et/ou édités dans telle conférence et/ou dans une année spécifique. Dans d'autres cas plus compliqués, le système se base sur des mots clés libres de choix, comme dans le cas des systèmes de recherche d'informations, le système renvoie une liste ordonnée d'articles, généralement, les *n* premiers qui sont les plus pertinents par rapport aux mots clés spécifiés. De même, la localisation basée sur la citation, comme le cas des moteurs de recherche basés sur la *citation* et *cocitation* (Bollacker, Lawrence, & Giles, 1998), qui par exemple, reçoivent un article en entrée et en se basant sur ses références renvoient une liste d'articles similaires du point de vue de leurs citations. Dans le deuxième cas de *localisation personnalisée*, le système implémente une perception des goûts personnels du chercheur qui représente son profil. Chaque usager a une vision subjective d'un article de recherche qu'il exprime sous forme d'évaluations suivant une échelle de valeurs bien déterminée. Cette expression de la valeur reflète une indication sur la qualité d'une ressource à travers une vue propre à cet usager. Le système, dit de recommandation, est capable d'interpréter ces évaluations, de les compiler et de prédire les ressources de qualité relatives à cet usager, dans le but de les lui proposer.

Un autre besoin des chercheurs est la possibilité de localiser des *parties spécifiques* d'un article (voir le paragraphe 3.1.1), permettant ainsi un niveau de *granularité* plus profond. Par exemple la possibilité de chercher un état de l'art relatif à un domaine, une

certaine problématique ou un article décrivant une telle expérience. La qualité d'un article n'est pas seulement sa valeur globale, mais elle dépend du *contexte* de recherche. Par exemple, pour un contexte de recherche qui est la localisation d'un article avec un bon état de l'art, la qualité contextuelle sera relative à ce critère. Donc, si un article a un bon état de l'art alors que, son évaluation globale est mauvaise, il sera considéré bon dans un tel contexte de recherche.

Il est aussi intéressant de permettre la recommandation non automatique de ressource entre les chercheurs et permettre le partage de connaissances. Cela peut être réalisé d'une façon explicite ou implicite. La première étant l'envoi explicite de l'information à un chercheur spécifique en utilisant, par exemple son courriel. Par contre, dans la façon implicite, un chercheur peut créer des liens entre des ressources jugées liées et documenter ce lien. Ces derniers seront, par la suite, mis à la disposition de la communauté des chercheurs qui pourrons les interpréter et les retracer facilement pour en bénéficier.

De nombreuses solutions dans la littérature scientifique, spécifiquement dans la recherche d'informations et le domaine des systèmes de recommandation, ont essayé de répondre à ces besoins. Nous allons présenter dans les sections prochaines les études jugées pertinentes et nous allons montrer leurs points forts et leurs points faibles dans le but de fournir une solution plus complète.

## 3.3 Les systèmes de gestion bibliographique

Les *systèmes de gestion de bibliographie* également désignés sous le nom de *systèmes de gestion de références* ou de *systèmes de gestion de citations*, sont des logiciels utilisés par les chercheurs et les auteurs, principalement pour enregistrer et gérer leurs citations ou références bibliographiques. Ces systèmes répondent, essentiellement, aux besoins de gestion bibliographique décrits ci-dessus. Nous allons présenter quelques études de cas pour illustrer ce type d'applications.

### 3.3.1 Principe de fonctionnement

En général, ces systèmes logiciels, dont l'architecture simplifiée est représentée dans la Figure 3.2, se composent d'une base de données, ou d'un *dépôt de références*, dans lesquels sont stockés tous les éléments composant la référence bibliographique, d'un dépôt de *styles* ou *formats de citations* qui contiennent les fichiers décrivant le style bibliographique.

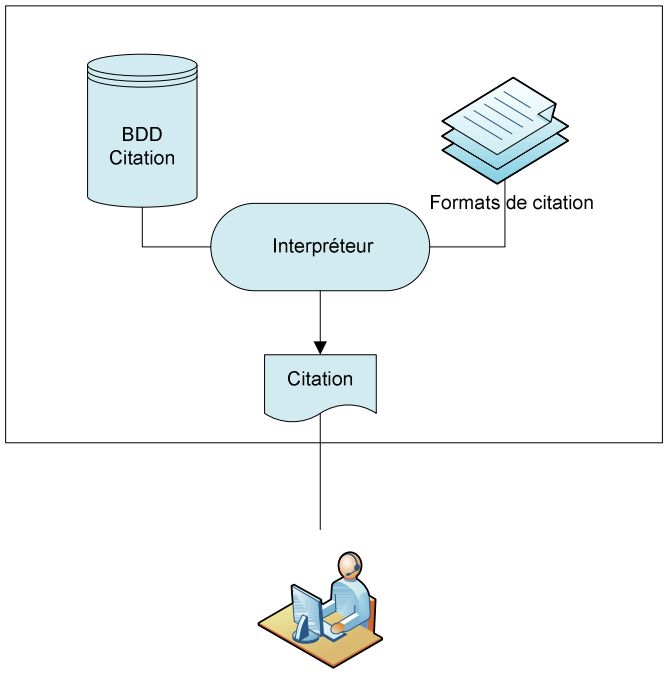



**Figure 3.2** Schéma simplifié d'un système de gestion de références

Et finalement, d'un *interpréteur* qui combine les éléments de la citation suivant le *style bibliographique* choisi par l'usager pour produire la *référence formatée*. Certains de ces systèmes fournissent un module de connexion avec les applications de traitement de texte via lequel elles sont capables de gérer directement la bibliographie du document. Cette fonctionnalité qui est appelée « *cite while writing* » décharge l'auteur de la tâche consistant à formater la bibliographie. C'est le système qui s'occupe automatiquement de l'ajout de la citation dans le texte à l'endroit voulu et insère sa référence correspondante dans la bibliographie à la fin du document, en veillant à respecter le format et l'ordre décrit dans le

style choisi. Le système intervient automatiquement, sur ordre de l'usager, pour changer quasi instantanément le style bibliographique.

Un système de gestion bibliographique de base, implémente des fonctionnalités élémentaires, telles que décrites ci-dessus, à savoir l'édition des éléments décrivant ses références, le formatage suivant un style donné, des outils auteurs pour la création et la modification de styles, auxquelles s'ajoutent les outils d'import et d'export de listes de références. Quelques-uns de ces systèmes sont dotés d'un module optionnel externe destiné à être inséré dans un éditeur de texte pour offrir des fonctionnalités qui permettent de manipuler ces références directement dans leur contexte d'insertion. Comme dans le cas d'EndNote (URL, 3), qui est une application basée localement sur l'ordinateur de l'usager, celle-ci offre la possibilité d'être intégré avec des unités de traitement de texte telle que Microsoft Word, sous forme d'un menu qui permet d'agir directement sur la bibliographie du document pour : formater, modifier, insérer ou supprimer des citations automatiquement.

La majorité de ces applications spécialisées dans la gestion de références sont étendues pour comprendre d'autres fonctionnalités supplémentaires relevant de la gestion de documents ou des systèmes de recommandation. Comme dans le cas de l'exemple cité ci-dessus, EndNote offre la possibilité de classer les références dans des répertoires virtuels. Le point faible de cette fonctionnalité est qu'elle n'offre qu'un seul niveau hiérarchique (pas de possibilité de créer des sous-répertoires). Les systèmes basés sur le web, comme CiteULike (URL, 4), permettent d'autres fonctionnalités, par exemple, de partager et de références au sein d'une même communauté ou d'un même groupe. CiteULike met aussi à la disposition des usagers internautes d'autres fonctionnalités dites de Web 2.0, comme l'ajout de commentaires, de tags, des revues et des évaluations (*rates*), qui à leur tour peuvent être partagés.

Les applications de gestion de références existent sous plusieurs formes. En effet, certaines sont basées localement sur l'ordinateur personnel de l'usager (par exemple EndNote), certaines sont basées sur le Web (comme CiteUlike) et sont accessibles via un navigateur ou client web, d'autres sont sous forme d'un module complémentaire (*plug-in)*

qu'un usager peut optionnellement ajouter à son navigateur (par exemple Zotero (URL, 11)). Nous présenterons chacune de ces applications dans les sections suivantes.

## 3.3.2 Études de cas d'applications de gestion de références

### 3.3.2.1 EndNote : un système basé localement

EndNote (Figure 3.3) est une application commerciale de gestion de références qui est produite par *Thomson Reuters*[1]. La version en cours (2009) est EndNote X2 pour les systèmes d'exploitation Microsoft Windows et Mac OS X.

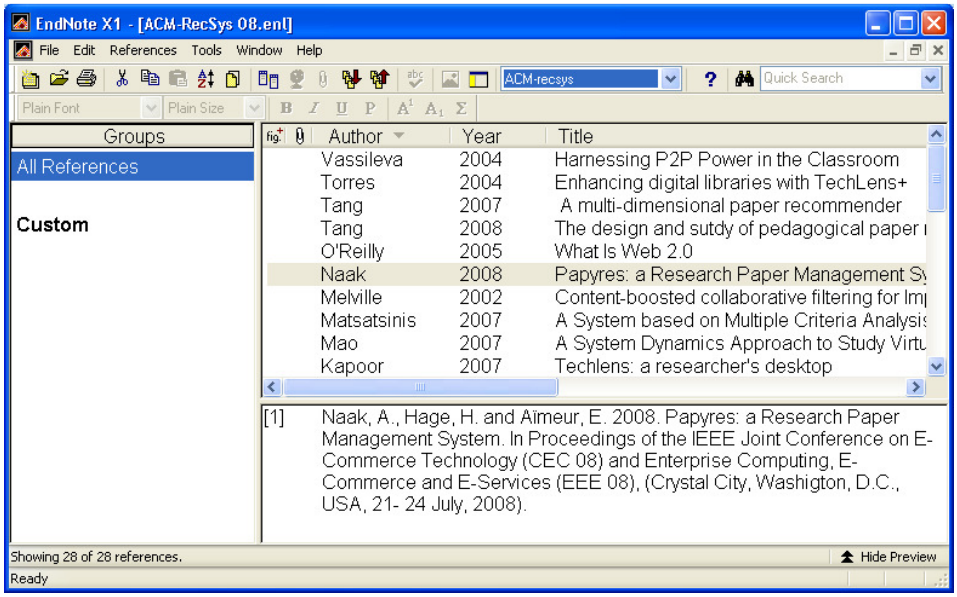



**Figure 3.3** EndNote

EndNote rassemble les références dans des fichiers de type « *EndNote library* » dont l'extension est (*.enl*) et peut éventuellement conserver une copie de l'article dans un dossier portant le même nom et se terminant toujours par (*.data*). Il offre plusieurs manières d'ajouter une référence à une bibliothèque : la saisie manuelle, l'importation à partir de certains types de fichiers, l'importation d'une autre librairie EndNote ou à travers une connexion vers des bibliothèques numériques via un protocole spécial. Chaque référence

[1] http://thomsonreuters.com/products_services/scientific/EndNote

enregistrée ou manipulée par cette application appartient à un ensemble de types prédéfinis, Nous citons entre autres, les types suivants : article de conférence, *article de journal*, *chapitre de livre*, *livre*. Il existe d'autres types comme : figure, table, film, image, etc., car cette application gère aussi d'autres sortes de références ou objets qui peuvent être insérées dans un document, par exemple : la table des matières et la liste de figures.

La saisie manuelle d'une référence commence par spécifier son type et le système affiche le formulaire correspondant et par la suite, complète les informations concernant les différèrent éléments de cette référence. Des champs s'étendent du général (auteur, titre, année) aux détails tels que le genre de référence (numéro d'ISBN, résumé, mots clés, temps de fonctionnement). Le système offre la possibilité d'attacher une copie physique de l'article à cette référence.

EndNote supporte plusieurs formats d'importation et d'exportation de références issus de plusieurs systèmes. La plupart des bases de données bibliographiques permettent à des utilisateurs d'exporter des références à leurs bibliothèques EndNote et vice versa. Ce moyen permet aux usagers l'insertion d'une multitude de références sans passer par une saisie manuelle une à une. Il prend en charge le protocole de communication inter bibliothèques, ce qui lui permet d'accéder à leurs bases de données via son interface pour rechercher dans leur catalogue les différentes ressources et importer leurs métadonnées correspondantes.

**Caractéristiques et fonctionnalités :**

- Connectivité avec les applications de traitement de texte : un module d'EndNote peut être installé sous Microsoft Windows qui apparaîtra dans le menu de Microsoft Word comme une barre d'outils.
- Prise en charge du protocole de connexion avec certaines bibliothèques. Il offre la possibilité de rechercher dans les catalogues de bibliothèque et les bases de données libres directement à partir de son interface.
- L'utilisateur peut saisir manuellement une référence en complétant ses champs correspondants.

- Permet d'importer une liste de références, dans ce cas, les champs correspondants seront automatiquement complétés.
- Une citation peut être automatiquement formatée dans près de deux mille modèles ou styles différents.
- EndNote peut exporter les bibliographies sous format texte, RTF (*Rich Text*), HTML ou XML.
- Pour chaque référence, EndNote permet de joindre son fichier correspondant sous plusieurs formats PDF (*Portable Document Format*) qui sera sauvegardé localement sur le disque et sera accessible à travers un lien.
- Offre la possibilité d'organiser les citations dans des répertoires à un seul niveau.

#### 3.3.2.2 CiteULike : un système basé sur le Web

CiteULike (Figure 3.4) est une application gratuite basée sur le Web. Elle offre un espace public où l'internaute peut chercher librement des articles (références) dans sa base publique. Il offre aussi un espace privé pour chaque usager qui s'enregistre en créant un compte avec un nom d'utilisateur et un mot de passe. Une fois dans son espace, l'usager peut commencer à ajouter ses propres références. CiteUlike est d'abord une application de collection de références qui ressemble à une bibliothèque numérique, telle que IEEE explorer (URL, 1) sauf que celle-ci est régie par ses membres participants. Elle permet d'ajouter manuellement des références, de les extraire automatiquement de certains sites et aussi d'exporter les listes de références sous certains formats tels que RIS (*Research Information Systems*), BibTex (format de fichier LaTex), RTF (*Rich Text Format*), et PDF. En revanche, il n'offre pas certaines fonctions de base comme le formatage des citations et des références suivant un style bibliographique défini, ni les outils auteurs de création de ces styles. En contrepartie, il intègre d'autres fonctionnalités qu'une application traditionnelle de gestion de références n'offre pas. En effet, cette application met à disposition de ses membres toute une plate-forme Web 2.0 pour contribuer, partager et manipuler des références, des articles et diverses informations. Par exemple, ajouter et

supprimer des références, télécharger en amont des copies numériques d'un article, les évaluer, les commenter, et faire leurs revues.

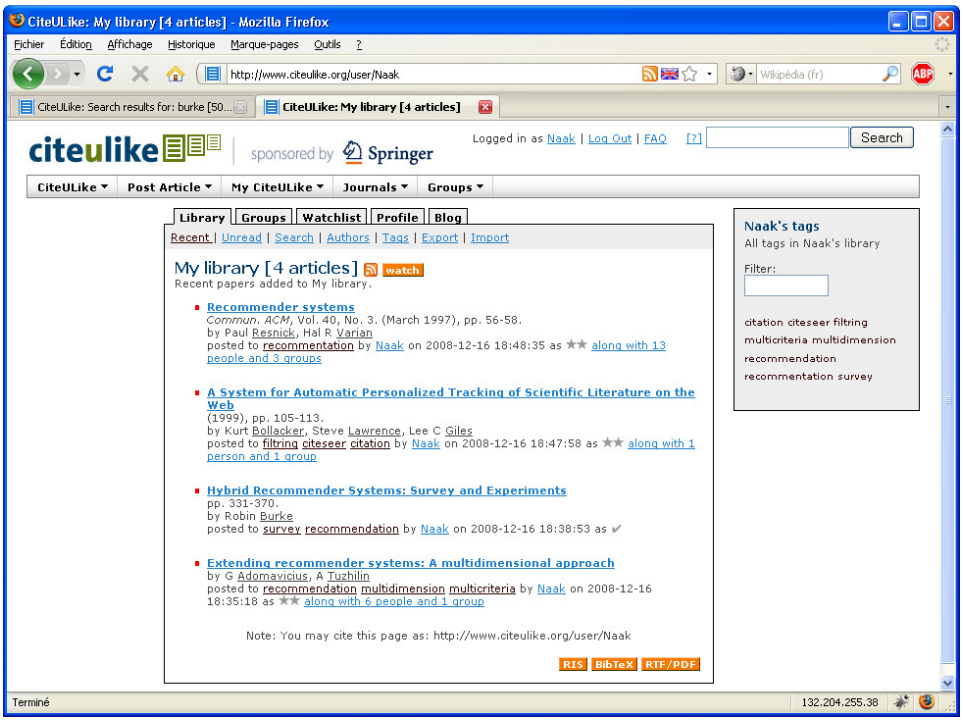



**Figure 3.4** CiteUlike : une application de gestion de références

Ajouter à cela la possibilité de créer un groupe ou d'adhérer à un groupe existant, de maintenir son propre blog, de participer aux forums. Il est basé sur le principe du marquage social (*social bookmarking*) et son objectif est de promouvoir et développer le partage des références scientifiques sur le web entre chercheurs. De la même manière que delicious[1] qui est capable de cataloguer des pages Web ou de Flickr[2] pour des photographies, les scientifiques peuvent partager toute sorte d'informations sur les articles de recherche.

**Caractéristiques et fonctionnalités :**

- Fonctionnalités de base pour la gestion de références : ajout manuel de citations, et extraction automatique de références à partir de certains sites Web.

[1] Son ancien nom est del.icio.us. Site web social permettant de sauvegarder et de partager ses marque-pages Internet. URL : http://delicious.com
[2] Un site web de partage de photos et de vidéos gratuit. URL : http://flickr.com/

- Import et export de listes bibliographiques sous plusieurs formats tels que texte brut, RIS, RTF, PDF, et BibTex.
- Fonctionnalités Web 2.0 pour les références : ajout de Tags, de commentaires, des revues, des évaluations, etc.
- Outils Web 2.0 pour la communication, notamment, les forums de discussion, des blogs, des listes de surveillance ou de suivi (*Watchlist*) et la diffusion de flux RSS.
- Contrôle d'accès : possibilité de créer des groupes ou adhérer à des groupes existants
- Partage de ressources telles que : les citations, les bibliographies et les articles de recherches
- Fonctionnalité « voisinage » qui peut être considérée comme une forme de recommandation élémentaire.

Malgré son qualificatif de système de gestion de références, CiteUlike ne permet pas la prise en charge du formatage de références suivant un style bien défini, ce qui est normalement une des fonctions de base de ces types de systèmes.

#### 3.3.2.3 Zotero : une extension pour Mozilla Firefox

Zotero (URL, 11) est une extension (*plugin*) gratuite et de code source ouvert (*open source*) qui propose un système de gestion de références destiné à être intégré au navigateur Mozilla Firefox. Elle a été créée par « *Center For history and New Media* » de l'université *George Mason*. Actuellement[1], il est à sa version stable 1.0.10 (Figure 3.5).

Comme tout système de gestion bibliographique, il permet la collecte, la gestion, et l'exportation de références sous différents formats. Intégrée au navigateur, lorsque ce dernier est sur un site web, Zotero est capable de reconnaitre les ressources littéraires, notamment les livres et les articles, et d'extraire les métadonnées correspondantes de ces pages en cours et de les sauvegarder dans un fichier local. Il peut éventuellement

[1] En date : juillet 2009.

sauvegarder une copie physique de ces ressources. Cette fonctionnalité d'extraction est surtout disponible pour les catalogues de bibliothèque et les librairies en ligne telles que Wikipédia[1].

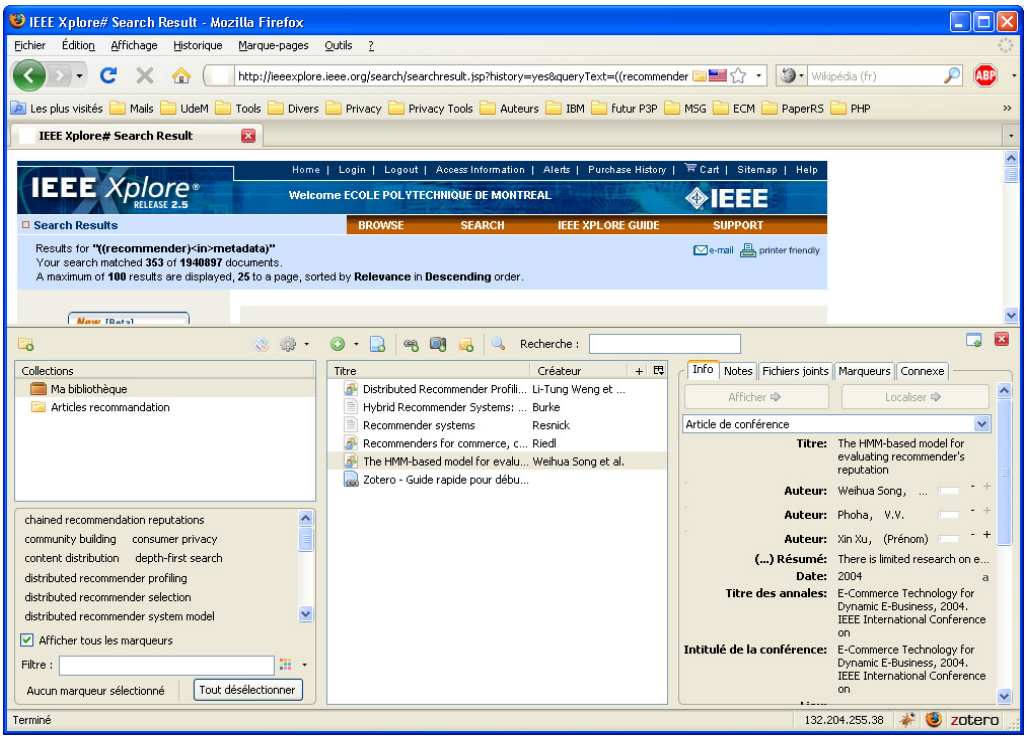



**Figure 3.5** Zotero : une extension pour Firefox

**Caractéristiques et fonctionnalités :**

- Gestion de base de références : ajout manuel de citation et extraction automatique de référence à partir de certains sites Web.
- Import et export de listes bibliographiques sous plusieurs formats
- Fonctionnalités Web 2.0 pour les références : ajout de Tags, de commentaires, des revues, des évaluations, etc.
- Fonctionnalité de gestion de documents : offre la possibilité de classer les ressources dans des dossiers.

[1] URL : http://fr.wikipedia.org/wiki/Accueil

- Fonctionnalité d'organisation et de gestion automatique des bibliographies
- Offre un module de connectivité avec les logiciels de traitement de texte.

### 3.3.3 Limite des systèmes de gestion bibliographique

Les applications de gestion de références sont d'une grande importance pour les chercheurs, rappelons qu'elles les déchargent des tâches routinières de citation et de référencement. La majorité de celles-ci ne se limitent pas à ces tâches, elles offrent d'autres fonctionnalités relevant de la gestion et de la localisation d'articles de recherche. En effet, le besoin d'autres fonctionnalités est ressenti sur plusieurs plans, c'est pour cela que plusieurs extensions ont été apportées, d'après ce que nous avons relevé des études de cas précédentes. Ces applications prises séparément ne répondent pas aux besoins de la recherche que nous avions recensés dans le paragraphe 3.2 et certaines de ces fonctionnalités de gestion, comme l'organisation en dossiers permet uniquement un seul niveau d'imbrication. En d'autres termes, les systèmes l'intégrant ne permettent pas la création de plusieurs niveaux de sous-répertoires. Le suivi de l'état d'un document est quasiment inexistant. Cela nous amène à conclure que la gestion de documents demeure une fonctionnalité superflue dans ce type d'application. Concernant la localisation de ressources, les fonctionnalités implémentées restent primitives, elles ne sont pas développées au même niveau que les applications dédiées pour la recommandation d'articles de recherche que nous allons voir plus loin dans ce chapitre. Pour bien cerner ces insuffisances liées à la gestion de documents et la localisation de ressources, nous allons étudier dans les deux paragraphes suivants les systèmes de gestion de contenu qui implémentent une gestion de documents très avancée. Par la suite, nous étudierons un autre genre d'application spécialisée dans la localisation d'articles de recherche, ce sont les *systèmes de recommandation d'articles de recherche*.

## 3.4 Les systèmes de gestion de contenu d'entreprise

Notre objectif de ce paragraphe est de donner une idée sur les systèmes de gestion de contenu d'entreprise et de voir à quel point ils pourront être utilisés dans la gestion

d'articles de recherche, du moins, s'inspirer des fonctionnalités de gestion de contenu pour traiter les insuffisances que nous avons mentionnées dans le paragraphe précédent concernant les systèmes de gestion de références.

### 3.4.1 Présentation

La gestion de contenu d'entreprise ECM (acronyme du terme anglo-saxon *Enterprise Content Management*) est un domaine très récent qui fait partie du concept générique « Gestion de contenu ». Ce dernier est très riche en facettes et englobe de son côté la gestion de contenu Web, la syndication de contenu, la gestion des ressources numériques et multimédias et, bien sûr, la gestion de contenu d'entreprise (ECM) (Kampffmeyer, 2006). Cette liste n'est pas exhaustive, car ce concept ne fait pas l'objet d'un consensus chez les acteurs de ce domaine. En effet, les concepteurs de ce type d'applications agissent en fonction de leur cahier des charges, de ce fait, ils dépendent des besoins du client. D'ailleurs, un rapport technique récent (Kampffmeyer, 2006), d'une firme de consultants spécialisée dans ce domaine, conclut : « Ce cercle vicieux des concepts indique surtout un manque de clarté dans les discours de marketing des producteurs. » Plus loin dans le rapport, le concept « ECM sera tout au plus, une vision, un synonyme d'une stratégie ou désignation d'un secteur commercial – mais certainement pas un système opérationnel ou un produit spécifique ». Pour les objectifs de notre étude, nous nous contenterons d'une définition englobant des fonctionnalités susceptibles d'améliorer la gestion d'articles de recherche.

### 3.4.2 Définitions

Les systèmes ECM traitent, plus particulièrement, des contenus non structurés. Parmi ceux-là, on retrouve tout genre de documents numériques ou numérisés : les courriels, les discussions enregistrées, les images, les pages Web, les fichiers multimédias, les articles de recherches, etc. Contrairement aux contenus structurés, ces contenus ne suivent pas une certaine logique qui pourrait faciliter leur sauvegarde et leur gestion comme dans le cas d'un document XML ou d'une base de données facile à accéder et à gérer (MARKESS, 2008). Pour cet effet, les ECMs, en plus de gérer des données structurées,

sont conçus pour faciliter le stockage, le partage, la gestion et la recherche d'informations au sein de ce type d'information non structuré, un défi auquel tout organisme ou entreprise fait face durant son existence.

Le concept d'ECM a évolué avec le temps et a été l'objet de plusieurs définitions à différentes périodes. En effet, en 2003, ECM a été défini par l'AIIM[1] (*Association for Information and Image Management International*) comme suit : "*The technologies used to capture, manage, store, deliver, and preserve information to support business processes.*" Et en 2005, ce concept a vu un autre changement avec la suppression de la notion de « processus » de la définition. Toutefois, l'AIIM a réaffirmé la gestion de processus métier BPM (*Business Process Management*) comme une composante essentielle. La définition devient : « *Enterprise Content Management (ECM) is the strategies, methods and tools used to capture, manage, store, preserve, and deliver content and documents related to organizational processes. ECM tools and strategies allow the management of an organization's unstructured information, wherever that information exists* ». La traduction, de cette définition enrichie est : « La gestion du contenu d'entreprise est l'ensemble des technologies, des instruments et des méthodes utilisés pour saisir, gérer/traiter, stocker, préserver et fournir des informations de soutien pour les processus commerciaux dans une entreprise » (Kampffmeyer, 2006). C'est cette définition qui est en cours d'utilisation, au moment[2] même de la rédaction de cette partie du présent mémoire.

### 3.4.3 Architecture générale

Les systèmes ECM sont, de plus en plus, des infrastructures logicielles très complexes. Non seulement pour l'explosion du volume de contenus survenue avec les nouvelles technologies de l'information, mais également, pour les nouvelles données économiques et légales contraignant les organismes à adopter de nouveaux règlements qui affectent le stockage et l'accès à certains types de contenu. L'architecture d'un ECM est complexe et

---

[1] http://www.aiim.org/

[2] Définition affichée sur le site web d'AIIM, dans la page Web : http://www.aiim.org/What-is-ECM-Enterprise-Content-Management.aspx# , accédée en juillet 2009.

diffère d'un produit à un autre, compte tenu de ce qui est décrit précédemment et aussi dépendamment des besoins et stratégies adoptées pour affronter cette complexité.

De manière générale, un système ECM, tel présenté dans la Figure 3.6, offre cinq catégories de fonctionnalités, à savoir : la capture, la gestion, le stockage, la distribution et la préservation à long terme ou l'archivage.

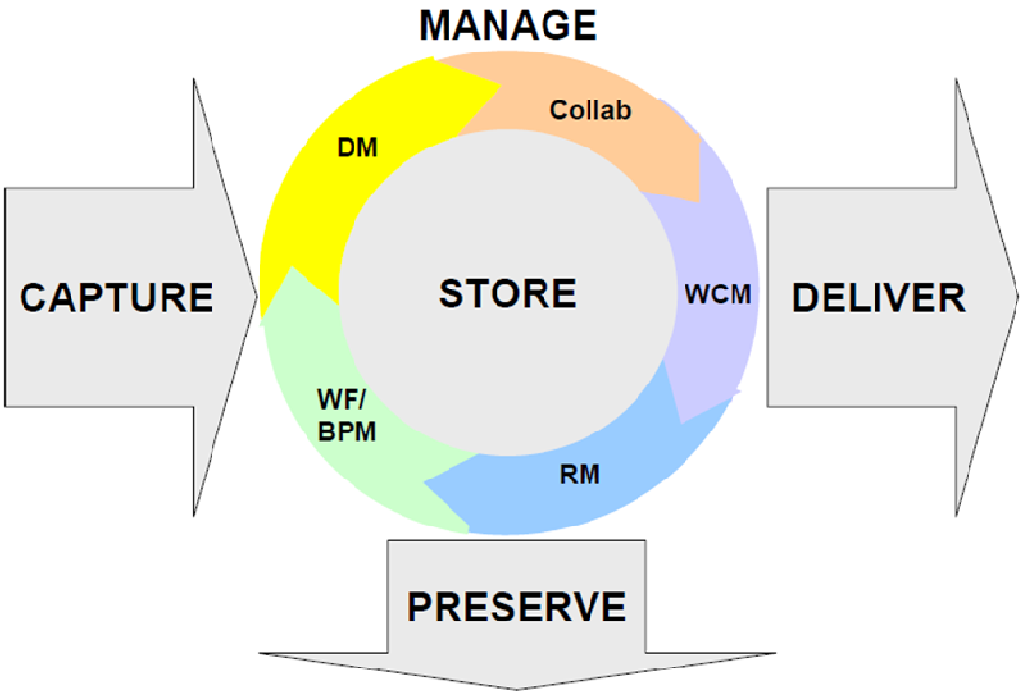



**Figure 3.6** Les cinq composants d'un système ECM (Kampffmeyer, 2006)

Dans le processus de gestion, on retrouve les cinq domaines connus dans la gestion de contenu, notamment :

- la gestion de documents DM (*Document management*),

- la collaboration Collab (*Collaboration*),
- la gestion de contenu Web WCM (*Web Content Management*),
- la gestion d'enregistrement RM (*Records Management*),
- *workflow* et gestion de processus d'affaire WF/BPM (*Workflow*/*Business Process Management*),

sous forme de composantes qui intègrent ces cinq catégories de fonctionnalités. Ces composantes sont des technologies autonomes qui peuvent fonctionner en combinaison ou d'une façon alternative (voir la Figure 3.6). D'autres technologies, telles que la gestion de messagerie, peuvent être associées à cette architecture, et inversement, certaines pourront être réduites. Tout cela pour dire qu'ECM est une agrégation de plusieurs technologies et sa force, justement, réside en cette combinaison.

**La capture (*CAPTURE*) :**

On retrouve dans cette catégorie « Capture » des fonctionnalités et des technologies pour la création, la saisie, le conditionnement et le traitement des informations analogiques et numériques. Cette catégorie fournit des moyens pour l'entrée de plusieurs genres de données (images, vidéos, documents, etc.) en utilisant différents moyens comme la saisie manuelle de document, la reconnaissance optique de caractère (OCR), la numérisation de documents, les données issues d'autre application telles que les applications de traitements de texte, les tableurs, la comptabilité, etc. L'objectif des technologies de capture est de rendre l'information sous une forme gérable à disposition du système ou de la composante de gestion pour des traitements approfondis ou pour l'archivage.

**La gestion (*MANAGE*) :**

Les composantes de gestion servent pour l'administration, l'élaboration et l'utilisation des informations. À cette fin, elles utilisent des

- Bases de données pour l'administration et le recouvrement ainsi que des
- Systèmes d'autorisation pour la protection d'accès et pour la protection d'informations.

On retrouve ici les sous composantes suivantes : la gestion de documents, la collaboration, la gestion de contenu Web WCM, la gestion d'enregistrement ou l'archivage RM, le *workflow* et la gestion des processus métier WF/BPM.

- **La gestion de documents DM** (*Document management*) : l'objectif de cette composante est de contrôler le cycle de vie du document, de sa création jusqu'à son archivage éventuel à long terme. Parmi les fonctionnalités qu'elle offre, on retrouve : la **Recherche et la navigation** qui facilitent la localisation de l'information. Le **Contrôle d'entrée et sortie** (*Checkin/Checkout*) pour contrôler la consistance des informations stockées. Le **Suivi des versions** et audit du document qui mémorise et contrôle l'évolution du document à travers le temps. La **Visualisation** comme l'organisation de l'information dans des structures telles que des dossiers et des sous-dossiers virtuels, des listes et des aperçus d'ensemble.
- **Collaboration (Collab)** comprend les fonctionnalités qui permettent aux membres d'un groupe de travailler et de coopérer pour la réalisation de tâches communes, à la manière des logiciels de groupe de travail (*Groupware*). On retrouve entre autres des moyens pour gérer les collections d'idées, la planification, et la gestion des projets. Aussi, d'autres moyens de communication issus de différentes technologies comme les vidéoconférences et des moyens issus du Web 2.0 peuvent aussi être utilisés.
- **La gestion de contenu Web, WCM** (*Web Content Management*) **:** cela inclut les portails (*portals*). L'intégration de cette composante facilitera la publication et la gestion de contenu mis en ligne (sur Internet, Extranet et les portails). Généralement, en séparant l'information de sa mise en forme moyennant différentes technologies comme XML (*Extensible Markup Language*), CSS (*Cascading Style Sheets*). La garantie de séparation de l'accès aux informations publiques et non publiques. Assurer les conversions et transformations nécessaires pour divers formats d'affichage, la mise à disposition et l'administration des informations pour la présentation Web.

- **La gestion d'enregistrement ou d'archivage** RM (*Record Management*) : Comprend aussi les systèmes de gestion de fichiers (*filing management systems*). Ce composant est responsable de l'administration de l'archivage et des sauvegardes régulières de l'entreprise. Des processus par lesquels l'entreprise sauvegarde d'une façon régulière et automatique pour une longue durée, mais limitée, des documents ou des informations importantes à conserver ou de nature obligeant à les conserver. Cela comprend l'administration de délais de conservation et des délais de destruction.
- **Workflow et Gestion de processus métier WF/BPM** (*Workflow/Business Process Management*) : Cette composante met à disposition des différents acteurs concernés par le processus métier les fonctionnalités et les moyens de gestion, de contrôle et de validation nécessaires à l'accomplissement de toute une chaine de production de ce processus. Pour un processus de publication en ligne par exemple, il s'agit de la modélisation des tâches de l'ensemble de la chaîne éditoriale.

**Le stockage (*STORE*):**

Cette composante concerne la sauvegarde temporaire des informations, elle est différente de la préservation qui est une sauvegarde à long terme. Par analogie à un bureau physique, la différence entre le stockage et la préservation est la même entre un dossier et une archive. Un composant de stockage typique est composé de dépôts comme un endroit de stockage (Base de données, Système de fichiers, *Data Warehouses*), d'un service de bibliothèque qui administre les zones de dépôt et les technologies de sauvegarde y compris les supports utilisés.

**Préservation et archivage (*PRESERVE*) :**

C'est la composante responsable de l'archivage à long terme de données ou d'informations importantes et qu'il est obligatoire de sauvegarder pour une certaine durée. Cette information est statique, stable et invariable, et n'est généralement pas utilisée, mais qu'il est obligatoire de sauvegarder. Par exemple le cas de données comptables des années antérieures. Ce composant se base sur différents supports (papier, microfilm, ...) pour

garder ces informations et utilise diverses techniques pour planifier, gérer et contrôler le bon déroulement de ses processus.

**Livraison et expédition (*DELIVER*) :**

Le rôle de ce composant est de produire des sorties d'informations en provenance des autres composants notamment : la gestion, le stockage et la préservation, et de les présenter sous différentes formes (fichier PDF, compressé, XML, syndication, ...). Ces sorties sont dirigées vers d'autres systèmes extérieurs, mais peuvent aussi être redirigées vers l'intérieur comme des entrées pour les mêmes composantes qui ont servi pour présenter leurs informations. Ce module comprend trois groupements de fonctions et de médias : technologie de transformation, technologie de sécurité et distribution. Les deux premiers groupes de fonctions sont disponibles pour toutes les composantes d'ECM de la même manière. La transformation regroupe des techniques de compression, de personnalisation, de transformation en différents formats (tels que PDF, XML). Les technologies de sécurité regroupent divers moyens comme la signature électronique, l'infrastructure PKI (Private Key Infrastructure) pour les clés, les certificats, l'intégrité, la prise en charge DRM (*Digital Rights Management*), la vie privée. La distribution regroupe tous les moyens de diffusion et de publication (Internet, web 2.0, support DVD, etc.).

### 3.4.4 L'article de recherche et les systèmes ECM

L'article de recherche et les métadonnées qui lui sont attachées (Figure 3.7) sont, du point de vue des systèmes ECM, composés de *contenus structurés* et *non structurés*. L'article de recherche, sous sa forme la plus élémentaire, est constitué de son contenu, généralement, en format PDF et de ses informations constituant sa référence. Le document texte est, de nature, un contenu non structuré. Nous avons eu l'idée de le décomposer en parties prédéfinies, lesquelles à leur tour sont des contenus non structurés. Quant à la référence, elle est un contenu structuré suivant une liste de métadonnées bien définies pour tout article de recherche. Cet article par la suite sera introduit dans un système de gestion, il se verra attacher des métadonnées de contexte qui dépendent de l'usager, tels que : les

*informations du contributeur* qui l'a introduit, la *date de la contribution*, l'*information de partage*, les *informations de suivi*. Ces informations sont de nature structurées.

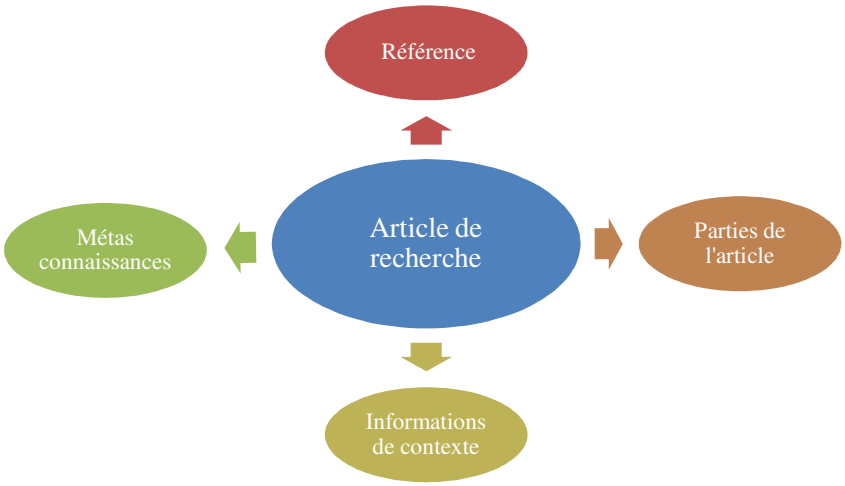



**Figure 3.7** Article de recherche : catégories de métadonnées

L'article évoluera dans le système et se verra attribuer, par son contributeur ou la communauté des chercheurs qui se le partagent, plusieurs méta-connaissances, telles que des commentaires, des résumés, des revues, des évaluations. Ces systèmes sont basés, généralement, sur le Web et offrent des outils dits Web 2.0, particulièrement pour cette fin. Ces méta-connaissances sont partagées entre structurées et non structurées.

Les ECM fournissent de nombreuses fonctionnalités intéressantes, notamment pour la gestion des articles de recherche, que les systèmes traditionnellement utilisés n'offrent pas ou offrent d'une façon primitive et élémentaire. Par contre, une telle complexité que celle que les ECM proposent de gérer n'est pas du même degré que celle d'un article de recherche. De ce fait, plusieurs fonctionnalités et composantes de cette architecture sont superflues. Et compte tenu du cas particulier des documents à gérer, en l'occurrence les articles de recherche, d'autres nouvelles fonctionnalités ou composantes spécialisées devraient être introduites pour ces besoins particuliers.

**Fonctionnalités de capture :**

La capture consiste en l'introduction de l'article de recherche dans le système ainsi que les informations qui l'entourent. À cet effet, les ECM offrent les moyens de saisie manuelle tels que les formulaires à remplir et des techniques avancées comme des dictaphones. D'autres moyens sont moins utiles dans notre contexte, par exemple le scanneur de document. Par contre, des fonctions spécialisées pour l'introduction des références ne sont pas disponibles, telles que les outils d'extraction automatique et d'import de référence qu'on trouve dans les systèmes de gestion de références. Cependant, il est utile qu'elles soient implémentées.

**Fonctionnalités de gestion :**

Sur ce plan, les ECM fournissent une plate forme très riche de fonctionnalités et de composantes pour l'administration, l'organisation et l'exploitation des informations du système. Nous les analyserons dans notre contexte en suivant les 5 composantes de l'architecture ci-dessus.

- Fonctionnalités de gestion de documents DM : on retrouve la recherche sous toutes ses formes, que ce soit dans les contenus structurés ou au sain même des contenus non structurés, notamment le texte de l'article et les métadonnées de connaissance attachées. Une autre catégorie de fonctionnalités est la *visualisation*, comme l'organisation des articles de recherches dans des dossiers et sous-dossiers virtuels, et l'affichage des informations de suivi des états de l'article tels que : lu, non lu, commenté ou pas. Le suivi de version sera plus utile si le système fournit des outils auteurs pour la production d'articles.
- Fonctionnalités de collaboration : comme dans le cas du suivi de version, la collaboration est utile dans le cas de la création ou de la production coopérative d'articles entre plusieurs chercheurs.
- Fonctionnalités de la gestion de contenu web : ces fonctionnalités sont obligatoires pour la gestion des transferts entre la partie privée des usagers et le domaine public. En effet, le système doit garantir la séparation entre données privées et publiques. Concernant la publication, si d'autres supports d'affichage sont utilisés le système devra adapter le format d'affichage en conséquence.

- L'archivage comme service complet et complexe tel qu'implémenté dans un système ECM n'est pas d'une grande utilité dans notre contexte. En effet, il n'y a pas de caractère obligeant à entretenir une stratégie d'archivage comme dans le cas d'une entreprise.
- Fonctionnalité de Workflow : dans une application de gestion d'articles de recherche, le processus de publication concerne, tout simplement, la mise en public d'informations que l'on a attachées en privé pour certains articles en les rendant publics. Ce n'est pas le cas, par exemple d'une chaine éditoriale d'un journal où la publication est un processus complexe où plusieurs acteurs interviennent à différents niveaux avec divers pouvoirs de décision.

**Fonctionnalités de stockage :**

Généralement le stockage des informations se fait dans des bases de données. Par contre, les copies physiques des articles sont sauvegardées localement, par exemple sur un disque dur. Le système utilise le système de gestion de fichiers fourni par le système d'exploitation du client.

**Fonctionnalités de livraison et d'expédition :**

Ces fonctionnalités dans un système ECM sont réparties suivant trois catégories, qui sont : la *transformation*, la *sécurité* et la *distribution*. Certaines *fonctionnalités de transformation* sont utiles pour un article de recherche et ses métadonnées. Notamment, la transformation en format PDF ou XML pour une exploitation interne par un autre composant du système, tel que la composante de gestion de contenu Web, pour par exemple personnaliser l'affichage des différentes métadonnées et réaliser des tris instantanés multicritère du côté client, en utilisant les différentes technologies de programmation. La transformation est aussi importante si le système compte utiliser les flux RSS et syndiquer des contenus. D'autres fonctionnalités de transformation, qui ne sont pas implémentées dans les ECM et qui concernent particulièrement les références, ce sont des fonctionnalités de formatage de ces dernières suivant un style donné. La deuxième catégorie de fonctionnalités de sortie concerne les *aspects de sécurité*. Il est intéressant d'implémenter un système de contrôle d'accès avec différents niveaux de permissions. Les fonctionnalités

de la sécurité sont aussi nécessaires telles que l'authentification et l'identification, la signature numérique, la communication sécurisée, et l'intégrité. Ce service pourra être étendu pour par exemple la gestion de droits électroniques DRM (*Digital Rights Management*) dans les échanges de copie physique d'articles. Il peut être étendu aussi pour des fonctions de transformation de données pour des fins de préservation de la vie privée. La troisième catégorie de fonctionnalités concerne *la diffusion et la distribution* de l'information sur divers supports. C'est la finalité où mènent les deux catégories précédentes, en effet l'information après avoir été transformée et adaptée suivant le média sera publiée ou diffusée par les sorties autorisées. Dans notre cas, pour une application web, les sorties sont l'affichage dans des pages web, les flux RSS, la messagerie électronique, les forums, l'enregistrement sur les supports usuels.

### 3.4.5 Exemples d'applications commerciales

Les systèmes ECM Livelink (URL, 10) et le système Documentum (URL, 12) qui sont respectivement les produits des compagnies *Open Text Corporation*[1] et *EMC2 Documentum*[2], sont parmi les systèmes utilisés dans les milieux d'entreprises pour la gestion de leurs contenus. Les deux solutions sont très riches en composantes et fonctionnalités, en plus des composantes de base présentées, ils s'étendent à d'autres, par exemple : la gestion de la messagerie, et la gestion de la connaissance. Les deux solutions évoluent suivant des versions au rythme des nouvelles technologies et standards pour soit les inclure ou alors s'y conformer.

Les ECM offrent de nombreuses fonctionnalités avancées pour la gestion de contenu dont l'article de recherche est un cas particulier qui peut bénéficier directement de leurs avantages. Mais ces systèmes restent inadaptés lorsqu'il s'agit de la gestion de références, un côté spécifique pour ce type de contenu. Plus encore, les besoins de localisation de ressources restent inassouvis. En effet, malgré les puissants moteurs de recherche dont

[1] http://www.opentext.fr/
[2] http://www.emc.com/

disposent ces applications, la recommandation demeure un service inexistant. Par conséquent, les systèmes ECM ne sont pas adaptés pour la gestion d'articles de recherche.

## 3.5 Les systèmes de recommandation d'articles de recherche

Avec la prolifération des conférences et des journaux, et le grossissement de l'éventail des travaux de recherche disponibles dans le Web, il devient de plus en plus dur pour les chercheurs de localiser des ressources qui répondent à leurs besoins. En effet, le nombre d'articles scientifiques publiés internationalement, dans les conférences par arbitrage (*peer-reviewed*), les journaux scientifiques et d'ingénieries, qui sont couverts par : SCI (Science Citation Index) et SSCI (Social Sciences Citation Index) avoisine les 700 000 en 2003 (Matsatsinis, Lakiotaki, & Delias, 2007). La base de données globale, toute discipline confondue en provenance de ISI Web of Knowledge1 atteint presque les 700 millions de références citées en 2008 (Reuters, 2008). Pour cela, diverses techniques de recommandation (Basu, Hirsh, Cohen, & Manning, 2001; Bollacker, Lawrence, & Giles, 1998; Miquel, Beatriz, pez, Josep, & s De La, 2003; Naak, Hage, & Aïmeur, 2009; Nishikant et al., 2007) sont utilisées pour aider les chercheurs à atteindre leurs objectifs de localisation des ressources appropriées.

### 3.5.1 Revues de littérature dans ce domaine

La recommandation d'articles scientifiques a suscité beaucoup d'intérêts dans la communauté des chercheurs. En effet, plusieurs études, expériences, prototypes et applications ont été réalisés pour aborder divers aspects et problématiques liées à ce sujet. Dans ce qui suit quelques revues de la littérature scientifique dans ce domaine.

#### 3.5.1.1 Techlens et les algorithmes de recommandation

Les auteurs, dans l'article (Torres, McNee, Abel, Konstan, & Riedl, 2004), ont expérimenté, à travers leur système TechLens, dix algorithmes de recommandation d'articles de recherche. Ces algorithmes sont basés sur le filtrage collaboratif, le filtrage à

[1] http://www.thomsonreuters.com/products_services/scientific/ISI_Web_of_Knowledge

base de contenu et les différentes combinaisons entre les deux dits filtrages hybrides. Ces algorithmes reçoivent en entrée un ensemble de citations[1] extraites automatiquement de la bibliographie d'un article cible (*active paper*) et produisent en sortie une liste ordonnée de citations comme recommandation. Les algorithmes basés sur le filtrage collaboratif se basent sur l'analyse des citations contenues dans l'article cible et génèrent en sortie une matrice « article-citation » en associant un vote pour l'article dans chacune de ces citations. Pour les algorithmes basés sur le filtrage à base de contenus, la technique de recherche d'informations TF-IDF (*Term Frequency/Inverse Document Frequency*) (Salton, 1989) est appliquée uniquement sur le titre et le résumé. Pour les algorithmes hybrides (Burke, 2002), les auteurs proposent « *feature augmentation*» et « *mixed recommenders* » qui combinent les deux précédentes. Leurs expériences ont été menées sur six catégories d'usagers : "*undergraduate student*", "*masters student*", "*PhD student*", "*researcher*", "*professor*" et "*professional"*. Les articles utilisés proviennent de la base de données publique CiteSeer[2] (URL, 8). Ces articles ont été répartis en 4 classes : "*novel*", "*authoritative*", "*introductory*", "*specialized*" et "*survey/overview*". Le système TechLens ne tient pas compte d'aspects de gestion concernant les articles de recherche. Du point de vue de la recommandation, il a été implémenté à la manière d'un moteur de recherche, au lieu d'utiliser directement des mots clés, il se sert de l'article d'où il extrait la liste des citations, le titre et le résumé pour servir de mots clés.

Récemment, les auteurs de l'article (Nishikant et al., 2007) ont proposé un système qui est basé sur TechLens, avec un aspect personnalisation et gestion de ressources. Le système opère dans un cadre privé et de groupe où plusieurs chercheurs peuvent collaborer pour un objectif commun. Dans le premier cas, l'usager pourra enrichir sa collection de citations avec la recommandation basée sur son profil. Dans un cadre de groupe, TechLens utilise la collection des différentes bibliographies des membres pour offrir des recommandations qui serviront leurs différents objectifs, tel que la proposition ou la recommandation d'articles pour la discussion ou pour la présentation au sein du groupe. Concernant l'aspect gestion, ce système permet la création et la gestion de groupes ainsi

[1] Le terme « citation » est utilisé indifféremment à la place de « référence » dans la littérature scientifique.
[2] CiteSeer est remplacé progressivement par CiteSeerX (*Next Generation CiteSeer*) (URL9)

que des fonctionnalités comme ajouter ou exclure des collèges, la possibilité d'ajouter des tags pour les articles, la possibilité de les rendre publics ou privés, et l'affichage d'informations utiles telles que les plus récents Tags et articles ajoutés. À ce stade encore, cette amélioration n'est pas suffisante, car plusieurs fonctionnalités sont absentes, par exemple tout l'aspect gestion des citations n'est pas abordé.

#### 3.5.1.2 Les systèmes avec approche pédagogique

Cette classe de système regroupe des applications développées dans un cadre éducationnel pour aider les étudiants ou les nouveaux chercheurs à localiser les ressources scientifiques dont ils ont besoin.

**Le système *Knowledge Sea II* (*KSII*)**

Knowledge Sea II (Peter Brusilovsky, Chavan, & Farzan, 2004; P. Brusilovsky, Farzan, & Jae-wook, 2005) est une plate-forme basée sur le Web et qui est utilisée pour étudier les moyens d'accès à l'information. Son objectif est d'aider les apprenants à découvrir d'autres sources d'informations, autres que leur support de cours habituels, tels que les tutoriels en ligne et les articles de recherche. Les auteurs ont utilisé les supports visuels et la navigation sociale qui sont basés sur le feedback explicite pour fournir quatre différents moyens d'accès à ces informations : la recherche, la carte d'information, la navigation hypertexte et la recommandation directe. La *recherche* est la fonctionnalité traditionnelle qui permet aux usagers de spécifier leurs mots clés et le système se charge de trouver les ressources qui les contiennent et les classe par ordre de fréquence. La *carte d'information* représente le noyau du système. Elle est composée de cellules qui contiennent des liens vers d'autres pages similaires. La *navigation hypertexte* est un autre moyen pour la localisation de l'information. En effet, les différentes leçons sont organisées sous forme d'une hiérarchie de liens ce qui permet de naviguer à travers les différents cours. Des liens directs peuvent être créés par les usagers pour lier la leçon et la ressource externe découverte par la recherche ou pendent la navigation dans la carte visuelle. *La recommandation* directe est utilisée par l'instructeur pour désigner les ressources importantes. Les liens correspondants à ces dernières seront ajoutés à la liste de ressource de la lecture correspondante dans le

portail des cours. Le système permet aux usagers d'annoter ou commenter les ressources pédagogiques. Ces dernières apparaissent avec des indicateurs visuels qui renseignent l'usager sur la présence de ces annotations. Des options intéressantes sont offertes, par exemple le partage et l'anonymat. Une annotation peut être privée ou partagée. De même, l'usager a le choix de s'identifier ou d'être anonyme.

Le système Knowledge Sea II est implémenté dans un cadre d'e-Learning, sa force réside dans les moyens divers qu'il met à disposition des apprenants pour localiser les ressources pédagogiques. Sur le plan gestion de l'information, le système n'est pas riche. Mis à part les deux fonctionnalités, qui sont le suivi de l'état de ressources et la possibilité de les lier, le système ne prend pas en considération les aspects propres à un article de recherche, par exemple la gestion des références et l'organisation des documents. Ce système les considère comme toutes autres ressources pédagogiques. Ajouter à cela, la recommandation telle que proposée par ce système est un feedback de nature sociale, alors que la recommandation basée sur les techniques personnalisés de filtrage d'informations n'est pas offerte.

**Système de recommandation pédagogique d'articles de recherche (Revue des travaux de Tang Tiffany)**

Les travaux de (Tang, 2008) s'inscrivent dans le domaine de la recommandation d'articles de recherche dans un contexte d'e-Learning « *pedagogical paper recommender systems* ». Elle démontre les spécificités pédagogiques de ce contexte dans ces publications (Tang & McCalla, 2004, 2007) et distingue plusieurs caractéristiques, dont les suivantes :

- Dans un contexte e-Learning, le système peut compter sur les annotations des usagers pour appliquer des recommandations à base de curriculum de cours. Une chose qui n'existe pas forcément dans un autre contexte.
- Dans un contexte pédagogique, la notion d'intérêt de l'apprenant doit tenir compte de son niveau et de son bagage intellectuel. Ainsi, les articles hautement évalués ne sont pas forcément ceux qui lui sont bénéfiques.

- La troisième différence concerne la qualité de la recommandation. Il ne suffit pas de mesurer la différence entre la valeur réelle et celle prédite, dans un milieu d'apprentissage. En effet, dans un tel contexte, la recommandation joue un rôle important pour offrir un support à l'apprenant et l'orienter. C'est pour cela que cette mesure doit prendre en considération la satisfaction des étudiants sur le plan pédagogique.

Après avoir mis en exergue les spécificités de ce contexte, divers algorithmes de recommandation pédagogique ont été implémentés en utilisant le filtrage à base de contenu, le filtrage collaboratif *multidimensionnel* (Tang & McCalla, 2007) et des techniques de filtrage à base de *modèles d'usager*. Les auteurs de cet article (Tang & McCalla, 2007) ont abordé la recommandation d'articles scientifiques d'un point de vue pédagogique. Dans un tel contexte, il faut tenir compte de certaines spécificités liées à ce domaine entre autres le niveau de l'apprenant, son bagage intellectuel. Ce contexte de recommandation est restreint et particulier comparativement à un contexte général et libre, notamment pour de nombreuses raisons telles que la supervision et l'assistance par un ou plusieurs tuteurs, le nombre limité d'usagers, les nouveautés et la diversité contrôlée des articles dans le système. Même le but est différent, il s'agit de maximiser les objectifs pédagogiques.

Ces travaux comportent des points intéressants par rapport aux objectifs de notre recherche, parmi les plus intéressants nous retrouvons la *recommandation multidimensionnelle* et la recommandation basée sur le profil des usagers. Sur le plan de la gestion d'articles de recherche, il n'y a que l'aspect localisation de ressources qui est abordé. Alors que les autres besoins, comme la gestion des références et la gestion de documents, ne sont pas considérés.

**Le système Comtella (Mao, Vassileva, & Grassmann, 2007; Vassileva, 2004, 2008)**

Comtella est un système de partage d'articles de recherche scientifique dans un milieu académique. À l'origine, il a été développé pour le partage d'articles de recherche entre les étudiants gradués du laboratoire de recherche. Par la suite, il a été généralisé pour comprendre plusieurs autres types de ressources pédagogiques. Ces articles, généralement en format PDF, sont téléchargés du Web et sont mis à la disposition des autres membres.

Le système offre aux étudiants la possibilité d'annoter des articles qui sont préalablement classés suivant leur sujet. Un usager voulant chercher une ressource devra d'abord spécifier un sujet et par la suite, reçoit une liste d'articles partagés et qui sont liés à cette catégorie. À partir de cette liste, la ressource désirée pourra être téléchargée, visualisée dans le navigateur, sauvegardée localement sous forme d'une copie. Les usagers ont aussi la possibilité de noter les articles partagés et d'y ajouter des commentaires ce qui peut aboutir à un classement global de ces articles en fonction de leur qualité et/ou de leur popularité au sein d'un groupe. C'est ainsi, une source supplémentaire d'information qui est générée automatiquement, et qui peut être très utile pour les nouveaux arrivants dans le laboratoire. Une version P2P (*peer to peer)* de Comtella (Wang & Vassileva, 2004) a été implémentée dans le but de pallier aux limites et défauts techniques de la version précédente et dans l'objectif d'étudier les problèmes liés à la motivation des membres d'une communauté virtuelle. Ce système est mis en application dans le contexte d'un cours. Basée sur un mécanisme incitatif, Comtella récompense les contributions des membres en utilisant un système hiérarchique (*gold, silver, bronze* et *common member*) basé sur le niveau de participation des usagers. Une participation active est jugée sur plusieurs points tels que : nombre de liens, meilleur lien, participation aux discussions, commentaires et évaluations.

Ce système implémente quelques fonctionnalités de gestion de contenu, dans sa version Web. En effet, il est capable de distinguer entre un domaine privé (personnel) et public (partagé) et d'adapter l'affichage suivant le cas. Il permet d'allouer les ressources nécessaires pour l'affichage de contenu, par exemple les fichiers de type PDF. Sur le plan de gestion de documents, il offre les fonctionnalités, ajout et suppression de ressources, partage, annotation et évaluation. Dans le domaine de la localisation, il offre uniquement la fonctionnalité de recherche. Ce système se distingue des autres par ses techniques de motivation. Cette application est restreinte au milieu académique, malgré les nombreuses fonctionnalités qu'il offre, plusieurs sont manquantes par rapport aux besoins que nous avions recensés, notamment la gestion de références et la recommandation d'articles de recherche.

#### 3.5.1.3 Les moteurs de recherche

D'une part, les systèmes tels que CiteSeer (URL, 8, 9) et Google Scholar (URL, 7) sont des moteurs spécialisés dans la recherche d'articles de recherche. Spécifiquement, CiteSeer indexe une gigantesque liste de références d'articles et est capable de fournir plusieurs statistiques telles que le nombre de fois qu'un article est cité, combien sont des autocitations. Google Scholar, quant à lui, analyse le texte des articles suivant les mots clés spécifiés au départ par un usager et il produit en sortie une liste d'articles de recherches classée suivant le poids de ces mots clés. Il donne aussi plusieurs informations, en liaison avec ces articles, telles que l'auteur, la publication, combien de fois l'article a été cité. Ces moteurs peuvent être considérés comme des systèmes de recommandation élémentaires qui fournissent des recommandations non personnelles. Certains articles proposent des systèmes de recommandation qui sont bâtis autour de ces derniers comme l'article des auteurs (Bollacker, Lawrence, & Giles, 1999). Ces derniers proposent un système de recommandation d'articles basée sur le moteur de recherche CiteSeer. Leur système se base sur des données de profils utilisateurs prédéfinis et hétérogènes. À l'aide de témoins (*cookies*), l'usager de cette application est suivi dans son activité pour analyser les mots clés qu'il utilise dans sa recherche afin de tenter de lui associer un des profils prédéfinis.

L'apport de ces systèmes se situe dans les besoins de localisation d'articles de recherche, cependant, ces systèmes ne prennent pas en considération la gestion de documents ainsi que la gestion des références.

#### 3.5.1.4 Autres

D'autres systèmes implémentés dans différents contextes sont proposés dans la littérature des systèmes de recommandation, la majorité d'entre eux sont des prototypes pour tester des algorithmes.

Les auteurs, dans l'article (Basu, Hirsh, Cohen, & Manning, 2001), se sont intéressés à la recommandation d'articles scientifiques dans le domaine de la revue d'articles soumis à des conférences avec arbitrage. Leur objectif est d'établir la relation entre un article et le profil d'un professeur susceptible de l'évaluer (*review*). Pour cela, ils

ont établi un modèle pour représenter l'article et un autre pour représenter le *reviewer*. Les articles sont représentés par leur titre, résumé et des mots clés bien choisis parmi une liste prédéfinie. Par contre, les informations sur le reviewer sont extraites automatiquement de ses propres articles. Les informations ainsi récoltées forment la matrice article-reviewer et pour obtenir une recommandation, il suffit d'interroger cette matrice. La recommandation dans ce contexte est différente de notre objectif qui est la localisation d'articles de recherches dans un but d'exploration et d'acquisition de connaissances.

Les systèmes, *Quickstep* et *Foxtrot* (Middleton, Shadbolt, & De Roure, 2004) utilisent un profil utilisateur *ontologique* pour la recommandation d'articles de recherche. L'approche utilisée consiste en trois phases. Dans la première, qui est hors connexion, le système classifie les articles suivant des critères respectant une certaine ontologie. La deuxième, en ligne, travaille en interaction avec l'usager. Le système demande à l'usager d'évaluer les articles avec les mentions : intéressant, non intéressant ou sans commentaire. Par la suite, le système capte automatiquement les réponses de l'usager pour constituer son profil. Enfin, dans la troisième partie, la recommandation est calculée par la corrélation d'articles précédemment explorés avec leur classification.

Dans l'article (Matsatsinis, Lakiotaki, & Delias, 2007), les auteurs proposent des algorithmes de recommandation multicritère d'articles de recherche qui mettent en œuvre une nouvelle approche, autre que celles conventionnelles en l'occurrence le filtrage collaboratif (CF), le filtrage à base de contenu (CBF) et hybride (voir chapitre 1). Cette approche est issue du domaine de la recherche opérationnelle et est basée sur les techniques MCDA (*Multiple-Criteria Decision Aiding*). Sur le plan de la recommandation, ces travaux sont hors du cadre défini dans cette présente recherche (voir paragraphe 2.2 ) et de plus, tous les autres aspects relatifs à la gestion ne sont pas pris en considération.

### 3.5.2 Récapitulatif

D'un point de vue localisation d'articles de recherche, tous les systèmes de recommandation d'articles de recherche précédents ne prennent pas en considération la *vision par parties* de l'article de recherche. Une vision importante qui accroit la flexibilité

de la recommandation et introduit un niveau de granularité plus fin avec la possibilité de localiser des sous-parties d'un article. En d'autres mots, il est intéressant pour un chercheur en quête par exemple, d'un bon état de l'art d'avoir des recommandations basées, principalement, sur ce critère. D'un point de vue algorithmique, aucun de ces systèmes ne prend en charge l'aspect multicritère. Par conséquent et pour répondre à cette contrainte, nous aurons à recourir aux approches exploitant le filtrage collaboratif multicritère que nous avions introduit comme préalable dans la section 2.5. Du point de vue de la gestion de documents et de la gestion de références, ces systèmes peuvent être considérés primitifs et sont loin de répondre aux exigences de la recherche.

## 3.6 Conclusion

Ce chapitre a mis en lumière les caractéristiques distinctives des articles de recherche auxquelles s'apparentent des besoins spécifiques dans l'activité de recherche. L'analyse de ces besoins nous a amenés à étudier trois classes de systèmes susceptibles de répondre à ces exigences, à savoir : les systèmes de gestion de références, les systèmes de gestion de contenu d'entreprise et les systèmes de recommandation d'articles de recherche. Tous ces systèmes pris séparément ne fournissent pas une solution satisfaisante pour la gestion et la localisation d'articles de recherche. Néanmoins, ce chapitre nous inspire de nombreuses fonctionnalités intéressantes que nous considérons pour répondre aux objectifs de cette recherche. En effet, dans le prochain chapitre, nous combinons les trois classes de système pour concevoir notre solution. Une attention particulière est accordée au moteur de recommandation afin d'explorer de nouvelles approches pour optimiser le choix du voisinage de prédiction utilisé dans le cadre du filtrage collaboratif multicritère.

# Chapitre 4 Conception de Papyres

Nous avons vu dans le chapitre précédent que chacun des types de systèmes, notamment les systèmes de gestion bibliographique, les systèmes de gestion de documents et les systèmes de recommandation d'articles de recherche, pris indépendamment, ne répondent pas aux besoins de gestion et de localisation d'articles de recherche ainsi que leurs métadonnées. Cependant, à notre connaissance, il n'existe pas de solutions globales qui répondent à ces besoins. Après avoir étudié et analysé ces trois types de système, nous nous en sommes inspirés pour concevoir Papyres. Dans ce qui suit, nous commencerons par une description générale du système, puis nous détaillerons le fonctionnement et la composition de ses différents modules.

## 4.1 Présentation générale

Papyres est une application de gestion et de recommandation d'articles de recherche avec tous les aspects qui leur sont liés. Ses fonctionnalités sont inspirées des trois systèmes étudiés dans l'état de l'art (Chapitre 3). C'est pour cela qu'il est positionné à la confluence de ces trois systèmes (Figure 4.1) : système de gestion de contenu d'entreprise, systèmes de gestion bibliographique et système de recommandation d'articles de recherche. Son objectif est de rassembler le maximum de fonctionnalités qui répondent aux besoins de la recherche analysés dans le chapitre précédent.

Papyres, opère sur deux plans : *privé* et *public*. Sur le plan privé, l'usager dispose d'un espace personnel où il a la possibilité de gérer ses références et ses documents et de localiser des ressources scientifiques. Sur le plan public, il offre la possibilité de partager des connaissances liées aux articles de recherche ainsi qu'un système de recommandation. Tout cela, dans une sorte de réseau social spécialisé qui offre plusieurs fonctionnalités Web 2.0 (O'Reilly, 2005) pour enrichir l'article avec de nouveaux contenus et pour améliorer la

communication et la diffusion. La combinaison de ces trois types de système constitue une des originalités de cette application.

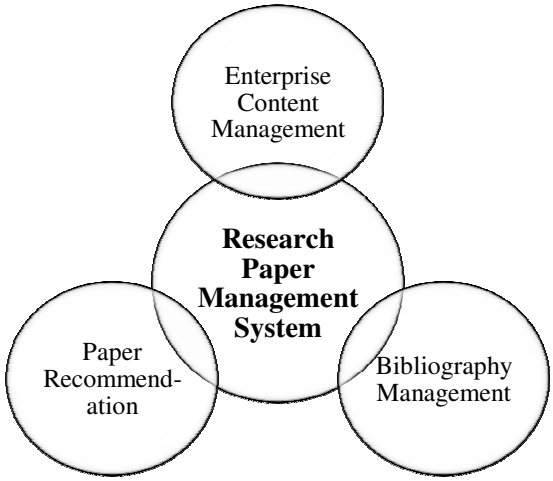



**Figure 4.1** Papyres: vue générale

L'utilisation de cette application requiert l'enregistrement de l'utilisateur sous un pseudonyme et un mot de passe, l'usager est ensuite invité à introduire quelques informations personnelles pour constituer son profil, entre autres son âge, son statut, ses intérêts de recherche. L'usager peut après cela ouvrir une session et accéder à son espace personnel où il dispose d'une multitude de possibilités pour utiliser le système.

L'espace personnel après sa création est vide, la première étape pour l'usager est de choisir ses articles de recherche. Cette opération est concrétisée par la saisie, dans un formulaire, des métadonnées de chaque article. En général, celles-ci sont représentées par les informations de la référence. Ces informations seront, ensuite, acheminées vers une base de données d'articles de recherche. Le système offre la possibilité d'attacher une copie numérique pour l'article introduit, celle-ci est enregistrée localement. Il offre aussi les moyens d'éditer son contenu dans un espace réservé à cet effet. Plusieurs fonctionnalités de base, comme la modification et la suppression, sont disponibles. Tous les articles ajoutés sont manipulés virtuellement par leur identifiant et non pas par leurs copies numériques. Ainsi, ils peuvent être organisés en dossiers virtuels imbriqués, être ajoutés aux favoris, être

partagés au sein d'un groupe ou être rendus publics, etc. La Figure 4.2 offre une vue générale des divers processus de Papyres.

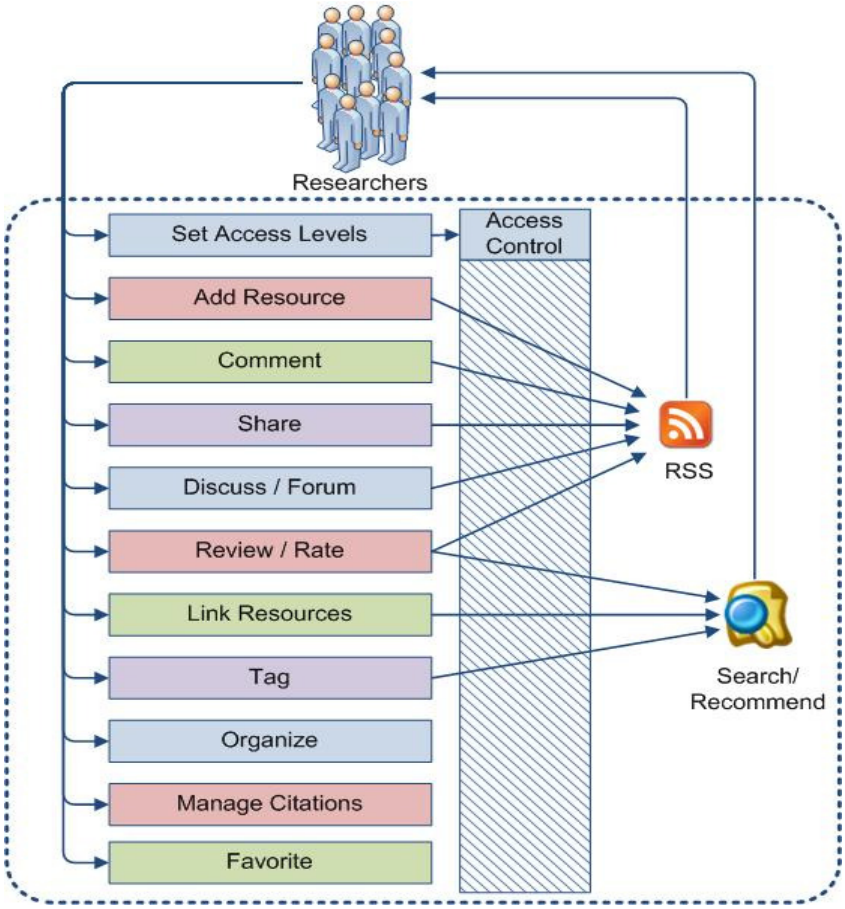



**Figure 4.2** Processus de Papyres

Le système offre aussi des fonctionnalités dites de web 2.0 pour l'enrichissement du contenu et l'amélioration de la diffusion et de la communication. En effet, l'usager peut créer du contenu et contribuer à enrichir les articles de recherche en ajoutant plusieurs méta-connaissances, comme les revues, les commentaires, les tags et les évaluations. Les flux RSS offrent un moyen efficace et personnalisable pour diffuser l'information. Les forums de discussion rendent possible l'établissement de communications asynchrones entre les différents usagers, sans les contraindre à être présents en même temps.

Un article, introduit dans le système, passe par plusieurs états durant son cycle de vie (lu ou non; document disponible ou non; appartient aux favoris ou non; commenté ou non; partagé ou non; etc.). Il est important d'avoir des fonctionnalités de suivi pour résumer son état sans dépenser temps et efforts.

Les références ou citations sont des aspects importants qui concernent l'article de recherche et compte tenu des objectifs de notre système, il est intéressant de mettre à la disposition du chercheur la possibilité de les gérer, de les formater avec différents styles bibliographiques et de les importer ou les exporter sous forme de listes, car cela épargnera l'effort répété de les organiser, à chaque fois que le contexte change.

La localisation de ressources est l'autre défi de Papyres, particulièrement, dans le domaine concernant la recommandation d'articles de recherche, les systèmes qui existent ne prennent pas en considération l'aspect multicritère relatif aux parties d'un article de recherche. Rappelons que ce dernier a été toujours considéré comme une entité indivisible. Dans un sondage (Naak, Hage, & Aïmeur, 2008) que nous avions mené, nous avons pu constater que fréquemment les chercheurs s'intéressent à certaines parties d'un article et ne lisent pas sa totalité et que cela dépend de leurs objectifs de recherche. Pour répondre à ce besoin, nous avons introduit une nouvelle vision pour l'article de recherche, en l'occurrence la vision par parties. En intégrant cet aspect dans notre système, nous avons pu arriver à un niveau de granularité plus fin en offrant par exemple la recherche et la recommandation de parties spécifiques d'un article.

Après avoir fait le tour de la majorité des fonctionnalités de Papyres, nous détaillerons chacun de ses modules dans ce qui suit.

## 4.2 La gestion de références dans Papyres

Tout article, introduit dans la base de données d'articles de recherche de Papyres, est caractérisé par des métadonnées (voir Figure 4.3), elles sont inspirées du standard LOM (*Learning Object Metadata*) (IEEE Learning Technology Standards Committee, 2002) et nous les avons classées suivant quatre catégories, *Général*, *Technique*, *Bibliographie* et *Contributeur*.

- Général (*General*) : cette classe comprend des informations générales sur l'article, comme son type (article de conférence ou de journal), ainsi que les différentes informations qui lui sont attachées, telles que les étiquettes (*tags*), les commentaires.
- Contributeur (*Contributor*) : cette classe regroupe les informations d'introduction de l'article dans le système, comme la date d'introduction et le nom de celui qui l'a introduit (le contributeur).

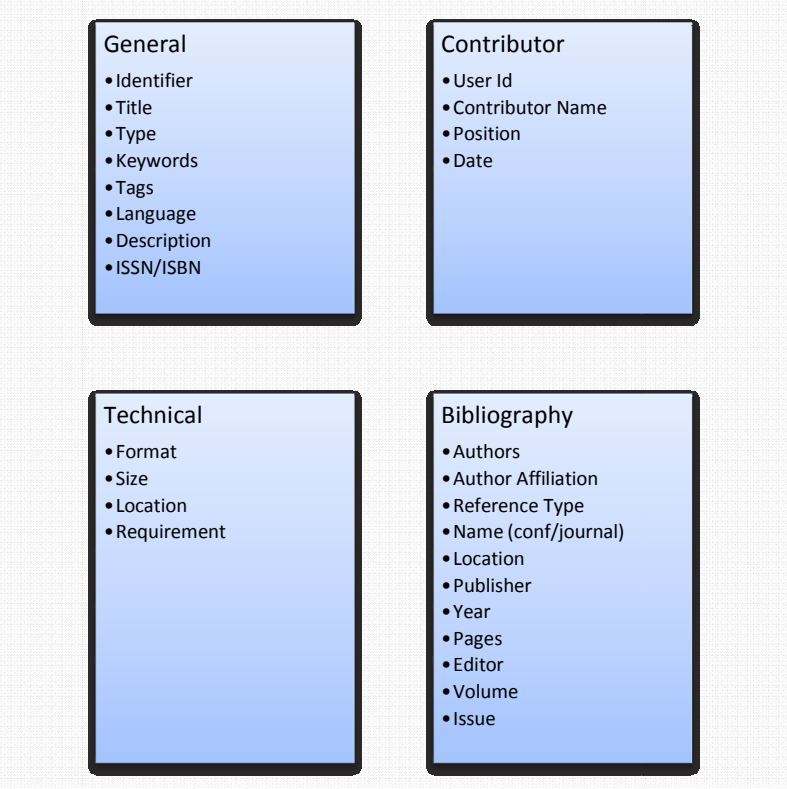



**Figure 4.3** Métadonnées de l'article de recherche.

- Technique (*Technical*) : cette classe contient des informations techniques concernant l'article introduit, comme le format de son contenu et sa taille.
- Bibliographie (*Bibliography*) : elle comprend les informations de citation, comme l'auteur ou les auteurs de l'article, et l'année de publication.

La référence d'un article de recherche est constituée d'un sous-ensemble de métadonnées appartenant aux catégories « Bibliographie » et « Général », qui est transcrit

suivant un style bibliographique donné. Trois concepts qui reviennent dans la gestion des références : la *référence*, la *citation* et le *style*.

- La *référence* d'un article de recherche est une représentation de ce dernier, qui est composée de métadonnées, au minimum, on y retrouve : le titre, le/les auteurs et l'année de publication.
- La *citation* est l'action d'évoquer la référence dans le texte de l'article qui renvoie vers la référence positionnée à la fin du document.
- Le *style* bibliographique décrit le format à respecter pour la citation de ressources et l'écriture de sa référence. On retrouve dans la description du style des informations sur les métadonnées qui entrent dans la référence ainsi que le format à respecter, tel que l'ordre des métadonnées et les signes de ponctuation à utiliser.

Chaque conférence fournit aux auteurs des directives à respecter pour formater leur article qui comprennent, entre autres, des informations sur le style bibliographique qu'ils doivent respecter, afin de faciliter le processus de publication.

### 4.2.1 Type d'articles de recherche

Suivant où l'article est publié, nous avons considéré trois types d'articles de recherche : *article de journal scientifique*, *article de conférence*, et *section de livre*. Concernant les articles de conférence, il n'y a pas de distinction, du point de vue de références et citations, entre un article long ou court (*Short paper*) :

- Article de journal : article publié dans un journal scientifique.
- Article de conférence : article publié dans une conférence scientifique, qu'il soit long ou court.
- Section de livre : c'est une partie d'un livre publiée sous forme d'un article dans une conférence ou un journal scientifique.

Même si la gestion de références va au-delà de ces trois types ci-dessus, le système est évolutif et pourra être étendu pour les prendre en charge. D'autres extensions futures

sont possibles notamment : la connexion des réseaux de bibliothèques, l'intégrabilité dans des logiciels de traitement de texte et un module complémentaire pour les navigateurs Web.

### 4.2.2 Style de la référence

Afin de faciliter la publication des actes de conférences (P*roceedings*) ou toute revue en général, le publicateur communique aux auteurs le format qu'ils doivent respecter pour présenter leurs travaux. Dans un appel de participation à une conférence, on le retrouve souvent dans la partie ou dans le document identifié par « *Call for paper* ». On y trouve plusieurs règles à respecter, parmi elles, le style des références. Ce dernier est identifié par un nom et il contient des informations sur le formatage de la référence. En général, ces informations indiquent les métadonnées de l'article qui vont entrer dans la construction de la référence, les signes de ponctuation à utiliser et la façon de les combiner.

Le module de gestion de références (Figure 4.4) exploite deux types d'informations qui sont les descriptifs de styles (le style, tout court) et les métadonnées de la référence qui sont respectivement localisés dans le *dépôt de styles* et la *BDD Articles* (base de données d'articles).

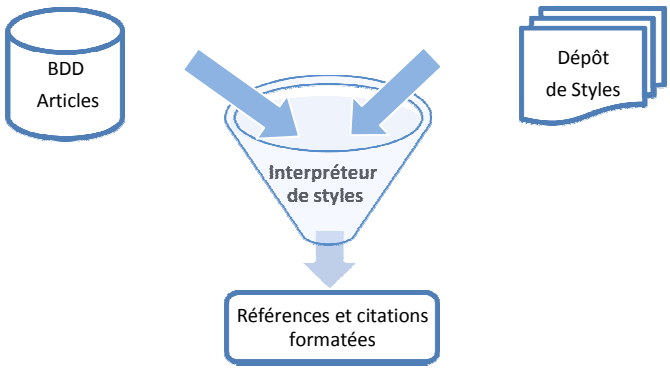



**Figure 4.4** Processus de formatage d'une référence

Au cœur de ce module se trouve l'*interpréteur de styles* dont le rôle est d'interpréter le style pour produire les *références et citations formatées*.

Le module de gestion de référence est en liaison étroite avec le module de gestion de documents. En effet, c'est à travers ce dernier que les différentes métadonnées d'articles sont disponibles dans le système. En général, ce module assure les différentes interactions entre le système et les usagers afin d'accomplir les diverses tâches de gestion et d'organisation de documents. Dans ce qui suit, nous présentons ce module.

## 4.3 La gestion de documents

On retrouve dans le système de Papyres les fonctionnalités de base d'un gestionnaire de documents, par exemple : ajouter, supprimer, modifier et partager un article, attacher, ou charger une copie de l'article dans le système, et afficher toute sorte d'informations liées à cet article.

### 4.3.1 Cycle de vie d'un Article dans Papyres

Un article de recherche dans le système Papyres est représenté par l'ensemble de ses métadonnées constituant sa référence et éventuellement, par son contenu qui peut être sous différents formats, comme : PDF (*Portable Document Format*), et DOC (Document Microsoft World). Lors de sa première introduction dans le système (Figure 4.5), le contributeur choisit s'il est *privé* ou *public*. Le choix privé peut être changé par la suite en le partageant, devenant ainsi public. Cette opération est irréversible, car une fois l'article est public, il est disponible pour tous les autres usagers de Papyres, et le fait de permettre l'opération inverse (redevenir privé), cela va créer des anomalies dans notre système. Dans l'espace privé, le système nous offre la possibilité d'organiser nos articles virtuellement. Cela veut dire que cette opération n'est pas physique. En d'autres termes, les articles demeurent dans un même emplacement sur le disque, en revanche, seulement leur identifiant est manipulé et classé dans des dossiers virtuels. Ainsi, l'article est enrichi d'informations sur le dossier et les sous-dossiers auxquels il appartient. La gestion d'articles avec Papyres nous offre la possibilité de choisir nos articles préférés et de les rendre accessibles dans un dossier appelé *favoris*. Il offre aussi une visualisation flexible de différentes informations, comme afficher les articles selon qu'ils appartiennent aux favoris ou selon leurs informations de partage (privés ou publics).

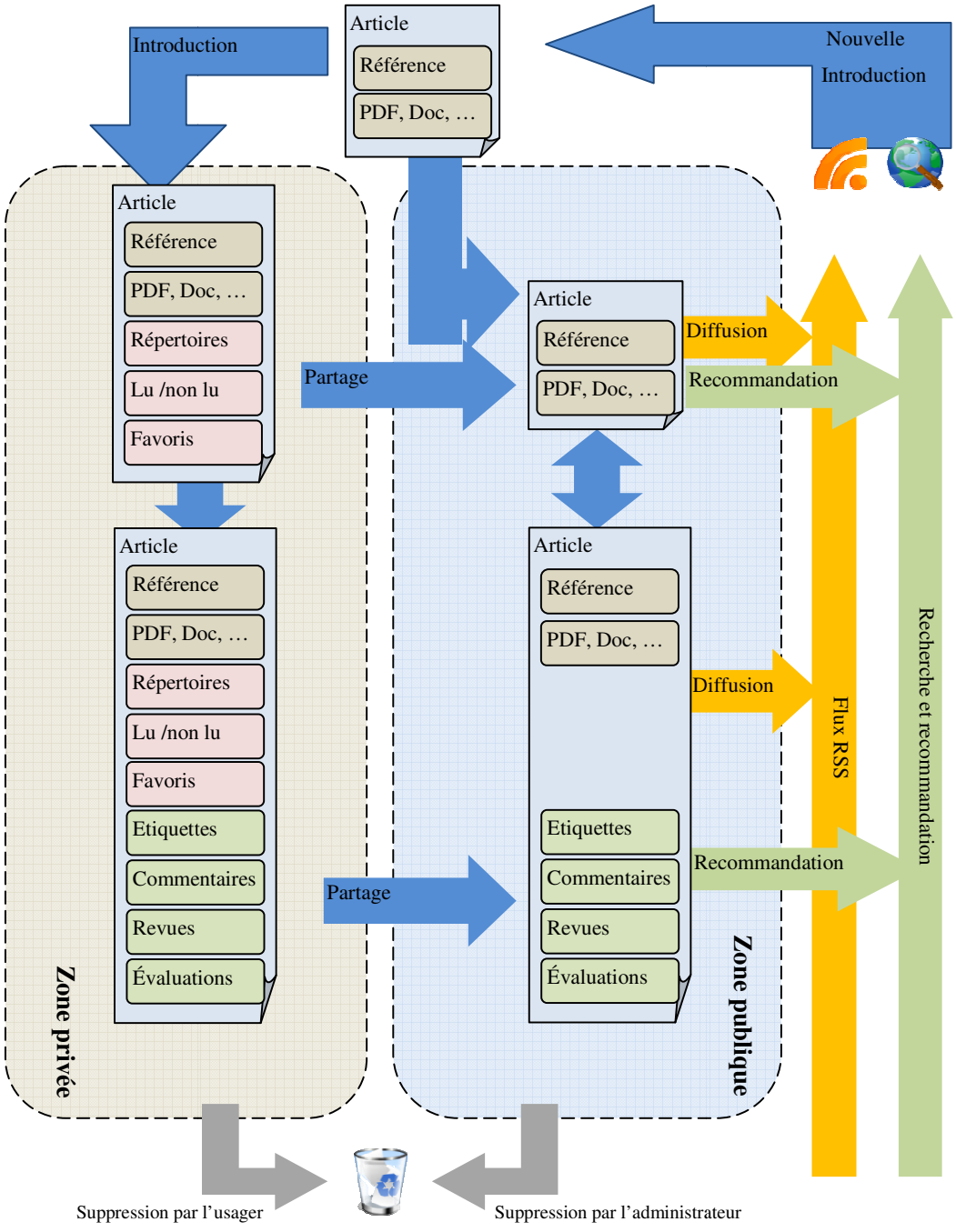



**Figure 4.5** Cycle de vie d'un article dans Papyres

Le suivi d'un article est une opération qui rassemble plusieurs états. Le premier étant, *lu* ou *non lu*, qui nous indique est-ce que nous avons ouvert le document ou pas. Les autres états concernent les métadonnées dites de Web 2.0 à savoir : l'article a-t-il été commenté, revu, évalué, ou étiqueté. Tout cela est indiqué pour chaque article, que ce soit dans l'espace privé ou public. Le fait de rendre public un article privé cela n'implique pas que toutes ses métadonnées vont devenir publiques. En effet, dans le cas des métadonnées personnelles qu'un usager a attachées aux articles, comme les commentaires et les revues, celles-ci sont privées par défaut et elles le restent ainsi tant que leur auteur ne les a pas rendues publiques explicitement. En effet, certaines informations ne seront pas partagées, telles que la façon dont les articles sont organisés en dossiers et sous-dossiers, les favoris et les informations de suivi ne le sont pas non plus.

Une fois l'article est dans l'espace public, il peut être diffusé par flux RSS. Lorsqu'il atteint un certain niveau d'évaluation, il pourra être recommandé. Il est même possible de le recommander explicitement pour un ami ou un collègue. Si l'article est accepté par un usagé, qu'il ait été reçu par flux RSS ou par recommandation, il sera introduit dans un autre espace privé, et l'usager pourra le classer à nouveau, le mettre dans ses favoris, ou encore commencer un nouveau suivi et le cycle est bouclé. Une fois l'article partagé, il demeurera ainsi, puisque cette opération est irréversible. Sauf dans le cas où l'article public est supprimé par l'administrateur, dans ce cas, l'article n'aura une nouvelle vie que si un contributeur le réintroduit dans le système.

La base de données d'articles de Papyres ne cesse de s'agrandir et les difficultés à repérer des ressources pertinentes augmentent en conséquence. Pour y remédier, nous avons implémenté le module recherche et recommandation, qui est l'objet de la prochaine section, afin de faciliter cette tâche.

## 4.4 Recherche et recommandation

Ce paragraphe explique le processus de localisation de ressources dans Papyres. Notons que nous distinguons deux types de localisations : la *recherche* proprement dite et la *recommandation*. Le premier type de localisation est basé sur l'usager, il spécifie lui-

même ses critères de recherche et par la suite, il analyse les résultats retournés. Dans le deuxième cas, la localisation de ressources se base sur les profils des usagers et elle est guidée par le système de recommandation.

### 4.4.1 La recherche dans Papyres

En général, un chercheur qui est motivé par une opération de recherche détient quelques informations caractéristiques de ce qu'il veut trouver. Ces informations initiales peuvent être classées en deux types : celles qui sont un sous-ensemble des métadonnées de l'article et celles qui sont un ensemble de mots clés relevant du sujet en question. Suivant le type de ces informations, nous distinguons deux types de recherche : la *recherche basée sur les métadonnées* et la *recherche basée sur les mots clés*. Dans le premier cas, nous parlons d'une simple recherche. L'usager spécifie ses critères qui sont une ou plusieurs des métadonnées de l'article, telles que l'auteur, la date. Puis le système envoie des requêtes vers la base de données d'articles pour retrouver les éléments correspondants. Dans le deuxième cas, la recherche est basée sur des mots clés spécifiés par l'usager. Pour trouver les articles qui sont pertinents par rapport à ces mots clés, des techniques de recherche d'informations sont nécessaires. Parmi ces techniques, nous retrouvons la mesure TF-IDF (*Term Frequency/Inverse Document Frequency*) (Salton, 1989). Le système analyse le texte de chaque article et lui associe un coefficient qui reflète sa pertinence par rapport aux mots clés spécifiés, puis il classe ces articles par ordre décroissant de ce coefficient. En d'autres termes, l'usager reçoit une liste ordonnée d'articles où les articles les plus pertinents se retrouvent en tête.

### 4.4.2 Le système de recommandation

Cette section détaille la méthode de recommandation utilisée dans le système Papyres. D'après la classification décrite dans le premier chapitre, notre approche de recommandation appartient à la classe des systèmes de recommandation hybride. Selon la classification de Burke (2002), c'est un hybride de type *cascade* (voir la section 2.3.3) où la première partie est un *filtrage à base de contenu* et la deuxième partie est un *filtrage collaboratif*. Concernant le filtrage collaboratif, selon la classification de (Adomavicius &

Tuzhilin, 2005), il est multicritère. En plus des approches de calcul de similarités proposées dans l'article (Adomavicius & Kwon, 2007), nous avons proposé quatre autres approches. Nous nous n'attarderons pas sur la première partie, car elle ne comporte pas de nouveauté. Par contre, nous accordons plus de détails pour la partie du filtrage collaboratif multicritère.

#### 4.4.2.1 Le filtrage à base de contenu CBF

Rappelons que la recommandation dans Papyres est de type Hybride. Sa première partie est un filtrage à base de contenu. Cette étape est très importante dans le processus de recommandation. En effet, grâce à elle, le système construit la liste d'articles représentant le contexte d'intérêt de l'usager. De même que la recherche basée sur les mots clés (voir le paragraphe précédent), le système retourne une certaine liste d'articles pertinents, cette fois-ci, cette liste servira comme une donnée d'entrée pour la deuxième partie du processus de recommandation. Cette partie du système de recommandation n'est qu'une préparation du domaine sur lequel un filtrage collaboratif sera appliqué. En termes de contribution, notre système ne fait qu'exploiter des techniques traditionnelles, dans ce domaine, et qui sont issues du domaine de la recherche d'informations.

#### 4.4.2.2 Le filtrage collaboratif multicritère

La recherche d'articles, comme présentée précédemment, ne considère que son aspect contenu. Cependant, l'aspect *qualité* est également un facteur important à considérer lors de la recherche. En effet, si un article de recherche peut satisfaire les exigences en matière de contenu, sa qualité peut ne pas satisfaire les attentes du chercheur ni répondre à ses besoins. Plus précisément, la qualité d'un document de recherche est relative. Elle n'est pas nécessairement reflétée par une « qualité globale ». En effet, comme déjà mentionné dans notre étude publiée (Naak, Hage, & Aïmeur, 2008), une grande majorité de répondants (composés d'étudiants en graduation et de professeurs) ont indiqué avoir souvent été intéressés par une partie spécifique d'un article, lors d'une opération de recherche de littérature. Cela nous a amenés à considérer cette partie qui est au centre d'intérêt du chercheur, car elle représente un facteur incontournable pour comprendre la qualité d'un article dans ce contexte relativement subjectif. Par exemple, considérons un chercheur

souhaitant explorer un nouveau domaine de recherche. Dans ce cas-ci, il sera plus intéressé par des articles de ce domaine avec un bon état de l'art, car ils ont une forte chance de satisfaire ses exigences de qualité, même si les autres parties de cet article ne sont pas pertinentes. Il est important de mettre à la disposition du chercheur des moyens pour localiser de nouvelle littérature, en précisant ses exigences de qualité pour une ou plusieurs parties d'un article. Nous désignerons cette qualité relative à une partie d'un article par les termes : *qualité contextuelle*. Nous verrons dans les sections suivantes comment que l'*évaluation multicritère* permet la prise en charge de ce nouvel aspect.

Afin de considérer le cas d'évaluations multicritère, il est nécessaire d'étendre les algorithmes de filtrage traditionnel pour prendre en charge l'aspect multicritère. Les algorithmes obtenus ainsi sont appelés *filtrage collaboratif multicritère*. Nous avons montré, dans la section 2.5 à l'aide d'un exemple détaillé, que ce type de filtrage multicritère est plus précis et pertinent qu'un simple filtrage monocritère. Nous avons aussi présenté quelques méthodes d'agrégation de similarités issues de la littérature scientifique, qui sont proposées dans le but d'adapter les formules traditionnelles de filtrage collaboratif monocritère afin qu'elles soient applicables dans le cas multicritère.

Le calcul de la similarité est une phase indispensable dans le filtrage collaboratif. Dans ce qui suit, nous commencerons par rappeler l'approche dite de similarité moyenne (Adomavicius & Kwon, 2007) utilisée pour trouver les $k$ meilleurs voisins relatifs à un usager cible. Par la suite, nous allons proposer quatre autres nouvelles approches.

#### 4.4.2.3 Critères d'évaluation d'articles de recherche

Les critères d'évaluation permettent aux chercheurs d'exprimer leur avis personnel concernant un article. Ces critères sont classés en trois types (voir Table 4.1) : les *critères généraux*, les *critères spécifiques des parties d'un article* et le *critère global*. Dans l'évaluation traditionnelle monocritère, nous ne retrouvons que le critère global qui nous renseigne à quel point le chercheur a aimé cet article. Alors que dans l'évaluation multicritère, en plus du critère global, les autres critères nous donnent le pourquoi de son choix. Ces évaluations sont un indice de qualité contextuelle qui est propre à un usager.

La classe des *critères généraux* nous renseigne sur l'article dans sa totalité. Elle est composée des quatre critères suivants : *Présentation*, *Orientation technique*, *Niveau technique* et C*lassification*.

**Table 4.1** Critères d'évaluation d'un article de recherche dans Papyres

| | | |
|---|---|---|
| **Critères généraux** | | |
| 1 | Présentation | Est-il facile à lire et à comprendre ? Est-il bien écrit et bien organisé ? Les idées, sont-elles claires ? |
| 2 | Orientation technique | Comment voyez-vous l'orientation technique de cet article ?<br>(1) Théorique, (2) Empirique, (3) Exploratoire,<br>(4) Pas d'avis, (5) Autre ? |
| 3 | Niveau Technique | Comment voyez-vous le niveau technique de cet article ? À quelle catégorie de chercheurs s'adresse-t-il ? Nécessite-t-il des pré-requis pour le comprendre ? |
| 4 | Classification | (1) Introductif, (2) Survol, (3) étude, (4) spécialisé,<br>(5) avancée |
| **Critères spécifiques aux parties de l'article** | | |
| 5 | Qualité de l'introduction | Le domaine, est-il bien introduit ? La problématique, est-elle claire ? |
| 6 | État de l'art | Est-il bien couvert ? Apporte-t-il un plus pour vos connaissances ? |
| 7 | Méthodologie | La méthodologie suivie, est-elle claire et convaincante ? |
| 8 | Expérimentation et validation | Y-a-t-il une validation des résultats dans cet article ? |
| 9 | Travaux futurs | Y-a-t-il une section pour les travaux futurs ? Y-a-t-il de bonnes perspectives de recherche ? |
| **Critère récapitulatif** | | |
| 10 | Évaluation globale | Globalement, trouvez-vous cet article intéressant ? |

Les *critères spécifiques des parties de l'article* permettent à l'usager d'évaluer différentes parties d'un article. Les parties de l'article considérées sont : *Qualité de l'Introduction*, *l'État de l'art*, *la Méthodologie*, *l'Expérimentation et validation*, et les *Travaux futurs*.

Le *critère récapitulatif* désigné par *Évaluation globale* résume l'avis général de l'usager vis-à-vis de cet article. C'est l'équivalent d'un rejet/acceptation dans la revue d'articles dans une conférence ou journal. Nous allons passer en revue chacun de ces critères en suivant l'ordre établi dans la Table 4.1.

- **Présentation :** L'usager donne son évaluation pour l'article du point de vue de la clarté de la rédaction et clarté des idées. Cette vue centrée sur la présentation de l'article nous renseigne sur le genre de présentation qui plaira à cet usager.

- **Orientation technique :** Un article peut avoir plusieurs orientations techniques, nous avons proposé de choisir entre : Théorique, Empirique, Exploratoire. L'usager peut ne pas spécifier ce critère ou spécifier un autre. Les choix possibles sont (1) Théorique, (2) Empirique, (3) Exploratoire, (4) Pas d'avis, (5) Autre.

- **Niveau technique :** ce critère nous renseigne sur la vision de l'usager concernant le niveau technique de l'article. Est-ce qu'il le perçoit de haut niveau, moyen, ou faible. Par exemple, un nouvel étudiant chercheur ne verra pas un article donné de la même perspective qu'un professeur. Les choix possibles sont : (1) Débutant (comme le cas d'un étudiant nouveau dans la recherche), (2) Débutant avancé, (3) Moyen, (4) Avancé, (5) Très avancé (comme le cas d'un professeur expérimenté).

- **Classification :** Ce critère donne l'occasion à un usager de classer l'article suivant cinq classes prédéterminées : (1) Introductif, (2) Survol, (3) étude, (4) spécialisé, (5) avancé.

- **Qualité de l'introduction :** l'introduction est la partie de l'article où est défini le domaine de la recherche. C'est aussi, la partie où l'auteur expose sa problématique et sa motivation. Cette partie peut contenir des définitions, des éléments de l'histoire du domaine, des références vers d'autres articles clés qui sont les piliers dans ce domaine. Un chercheur, même très avancé, peut avoir besoin d'explorer de nouveaux domaines de recherche et trouver cette partie de l'article très intéressante.

- **Qualité de l'état de l'art :** On ne peut pas convaincre la communauté scientifique de l'originalité d'un travail de recherche et de sa contribution si l'état de l'art n'est pas bien fait ou est incomplet. Cette partie de l'article cible les travaux qui lui sont similaires pour montrer la force de son apport et se distinguer et se démarquer. Un bon état de l'art est l'équivalent d'une mine d'or pour un chercheur en train de rédiger la même partie pour son article du même domaine. En effet, il économisera temps et effort pour trouver les articles similaires.

- **Méthodologie :** la méthodologie suivie dans l'article est très importante. C'est là que se trouve l'approche et la méthode entreprise par l'auteur pour analyser et résoudre la problématique. Par exemple, si cet article est fait par un auteur expérimenté, cette partie de l'article sera une source très intéressante d'apprentissage pour un débutant. Elle peut être aussi une inspiration pour la résolution de problèmes similaires.

- **Expérimentation et validation :** cette partie peut contenir plusieurs informations précieuses par exemple : la métrique utilisée, la méthodologie de l'expérience, une méthode d'analyse. Donc, cette partie pourra être une cible directe pour des chercheurs qui ont les mêmes objectifs.

- **Travaux futurs :** un des indices de la portée d'une recherche et l'étendue de son horizon. La discussion des travaux futurs dans un article peut combler plusieurs lacunes et faiblesses. En effet, si un auteur ne couvre pas certains côtés dans sa recherche, ces derniers pourront être une cible de critiques pour de nombreuses raisons par exemple : le cadre limité de cette recherche est court. Le fait de les aborder comme travaux futurs est une façon de se couvrir. Cette partie, et une ouverture d'horizon et une inspiration pour par exemple de nouveaux étudiants en recherche d'un sujet, d'étude de cas ou de travaux pratiques.

- **Évaluation globale :** ce critère ne représente pas forcément une moyenne des évaluations précédentes, mais il est évident qu'il est lié aux autres critères. Plus précis, il y a corrélation entre lui et les autres critères. Comme nous l'avons dit

précédemment, c'est l'équivalent d'un rejet/acceptation dans la revue d'articles d'une conférence ou d'un journal.

### 4.4.3 Échelle d'évaluation

Nous avons choisi une échelle variant de 1 à 5 pour l'évaluation de ces critères. Le chiffre 1 étant la plus haute expression du refus (rejet) et le chiffre 5 est la plus haute appréciation (acceptation). Le choix d'un nombre impair permet un choix intermédiaire qui représente l'état neutre ou indécis. L'étendue relativement courte (5 choix) facilite à l'usager d'exprimer son avis sans ambigüités. Car, étaler les choix sur une longue échelle, par exemple de 1 à 17, peut entrainer une perte de nuances. En effet, les valeurs rapprochées ne sont pas déterministes. Dans l'exemple précédent, les valeurs 15 et 16 présentent peu de différence. D'où l'importance de choisir une échelle d'évaluation appropriée (Herlocker, Konstan, Terveen, & Riedl, 2004).

### 4.4.4 Approches pour trouver le voisinage

Avec le filtrage collaboratif, il est impératif de trouver le voisinage de l'usager cible auquel est destinée cette recommandation. La qualité de celle-ci est fortement liée à la qualité de ce voisinage. Nous avons abordé, dans le chapitre 2 concernant les systèmes de recommandation, le filtrage collaboratif multicritère comme proposé dans l'article (Adomavicius & Kwon, 2007). Nous commencerons par reprendre l'approche « moyenne des similarités » que nous désignons dans ce présent mémoire par **HZ** (*HoriZontal*) afin de la tester et la comparer avec nos quatre nouvelles approches en l'occurrence : **VL** (*VerticaL*), **HZ-VL** (*HoriZontal* then *VerticaL*), **VL-HZ** (VerticaL then *HoriZontal*) et finalement, horizontale sans bruit **HZ-N** (*HoriZontal* without *Noise*). Notons que la différence entre ces méthodes réside dans la manière de choisir ce voisinage à utiliser dans la prédiction. Nous étudions en détail chacune de ces approches.

**Similarité de Person pour un critère :**

Soit : $U = \{u_1, u_2, \dots u_n\}$ *ensemble de n usagers*

$S = \{s_1, s_2, \dots s_m\}$ *ensemble de m items*

Chaque item $s_i$ est évalué suivant $l$ critères $c_j$.

Soit :

$C = \{c_1, c_2, \dots c_l\}$ *ensemble de ces l critères*. Dans notre cas : $l = 10$ critères.

La similarité $sim_{c_i}(u, u')$ entre un usager $u$ et un usager $u'$ par rapport à un critère $c_i$ est donnée par la formule de corrélation de Person :

$$sim_{c_i}(u, u') = \frac{\sum_j (v_{uj} - \bar{v}_u)(v_{u'j} - \bar{v}_{u'})}{\sqrt{\sum_j (v_{uj} - \bar{v}_u)^2 \sum_j (v_{u'j} - \bar{v}_{u'})^2}} \quad (1)$$

$j$ : Nombre d'objets ayant été votés à la fois par $u$ et $u'$

$v_{uj}$ : Vote de $u$ pour le critère $j$

$\bar{v}_u$ : Moyenne des votes de $u$ pour un critère $c_j$

**Illustration pour 5 usagers, 4 items et 4 critères :**

La Table 4.2 illustre cinq usagers qui évaluent quatre items, le résultat est une matrice de 20 éléments (évaluations).

**Table 4.2** Exemple d'une matrice d'évaluations classique *Usagers x Items*

| | **$s_1$** | **$s_2$** | **$s_3$** | **$s_4$** |
|---|---|---|---|---|
| **$u_1$** | $v_{11}$ | $v_{12}$ | $v_{13}$ | $v_{14}$ |
| **$u_2$** | $v_{21}$ | $v_{22}$ | $v_{23}$ | $v_{24}$ |
| **$u_3$** | $v_{31}$ | $v_{32}$ | $v_{33}$ | $v_{34}$ |
| **$u_4$** | $v_{41}$ | $v_{42}$ | $v_{43}$ | $v_{44}$ |
| **$u_5$** | $v_{51}$ | $v_{52}$ | $v_{53}$ | $v_{54}$ |

Cette fois-ci, chaque élément $s_i$ ($i$ varie de 1 à 4) possède quatre critères $c_l$ ($l$ varie de 1 à 4), la matrice précédente se transformera comme illustré dans la Table 4.3 et son nombre d'éléments s'agrandit à 80 évaluations

**Table 4.3** Exemple d'une matrice d'évaluations multicritère *Usagers x Items*

| | $S_1$ | | | | $S_2$ | | | | $S_3$ | | | | $S_4$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathbf{c_1}$ | $\mathbf{c_2}$ | $\mathbf{c_3}$ | $\mathbf{c_4}$ | $\mathbf{c_1}$ | $\mathbf{c_2}$ | $\mathbf{c_3}$ | $\mathbf{c_4}$ | $\mathbf{c_1}$ | $\mathbf{c_2}$ | $\mathbf{c_3}$ | $\mathbf{c_4}$ | $\mathbf{c_1}$ | $\mathbf{c_2}$ | $\mathbf{c_3}$ | $\mathbf{c_4}$ |
| $\mathbf{u_1}$ | $v_{111}$ | $v_{112}$ | $v_{113}$ | $v_{114}$ | $v_{121}$ | $v_{122}$ | $v_{123}$ | $v_{124}$ | $v_{131}$ | $v_{132}$ | $v_{133}$ | $v_{134}$ | $v_{141}$ | $v_{142}$ | $v_{143}$ | $v_{144}$ |
| $\mathbf{u_2}$ | $v_{211}$ | $v_{212}$ | $v_{213}$ | $v_{214}$ | $v_{221}$ | $v_{222}$ | $v_{223}$ | $v_{224}$ | $v_{231}$ | $v_{232}$ | $v_{233}$ | $v_{234}$ | $v_{241}$ | $v_{242}$ | $v_{243}$ | $v_{244}$ |
| $\mathbf{u_3}$ | $v_{311}$ | $v_{312}$ | $v_{313}$ | $v_{314}$ | $v_{321}$ | $v_{322}$ | $v_{323}$ | $v_{324}$ | $v_{331}$ | $v_{332}$ | $v_{333}$ | $v_{334}$ | $v_{341}$ | $v_{342}$ | $v_{343}$ | $v_{344}$ |
| $\mathbf{u_4}$ | $v_{411}$ | $v_{412}$ | $v_{413}$ | $v_{414}$ | $v_{421}$ | $v_{422}$ | $v_{423}$ | $v_{424}$ | $v_{431}$ | $v_{432}$ | $v_{433}$ | $v_{434}$ | $v_{441}$ | $v_{442}$ | $v_{443}$ | $v_{444}$ |
| $\mathbf{u_5}$ | $v_{511}$ | $v_{512}$ | $v_{513}$ | $v_{514}$ | $v_{521}$ | $v_{522}$ | $v_{523}$ | $v_{524}$ | $v_{531}$ | $v_{532}$ | $v_{533}$ | $v_{534}$ | $v_{541}$ | $v_{542}$ | $v_{543}$ | $v_{544}$ |

**Prédiction :**

$k_{voisins}(u) : ensemble\ des\ k\ voisins\ les\ plus\ similaires$ à $l'usager\ cible\ u$

$P_{u\,c_i}$: $prediction\ de\ u\ pour\ le\ critère\ c_i$

$$P_{u\,c_i} = \bar{v}_u + \frac{\sum_{j=1}^{k} sim_{c_i}(u, u_j)\,(v_{jc_i} - \bar{v}_j)}{\sum_{j=1}^{k} sim_{c_i}(u, u_j)} \qquad (2)$$

$u_j \in k_{voisins}(u)$

#### 4.4.4.1 Rappel de l'approche moyenne des similarités (HZ)

Cette approche correspond à la moyenne des similarités (*average similarity)* (Adomavicius & Kwon, 2007), et appartient à la classe des méthodes désignées par « *Aggregating traditional similarities from individual criteria* » dont le but est d'étendre les algorithmes de filtrage collaboratif traditionnel basé sur un seul critère pour s'adapter au multicritère. Après avoir calculé les similarités individuelles $sim_i(u, u')$, nous calculons la moyenne de ses similarités $sim_{avg}(u, u')$ qui est donnée par la formule suivante :

$$sim_{avg}(u, u') = \frac{1}{l} \sum_{i=1}^{l} sim_{c_i}(u, u') \qquad (3)$$

**Illustration pour 4 critères et 5 usagers :**

Le nom HZ (*HoriZontal*) vient de la Table 4.4 ci-dessous. Trouver la moyenne des similarités, entre l'usager cible et chacun des autres usagers par rapport à chacun des critères, revient à calculer la moyenne suivant les lignes horizontales de cette table.

**Table 4.4** Matrice de similarités et approche HZ

| **Paire :** | $c_1$ | $c_2$ | $c_3$ | $c_4$ | ***Sim* Moyenne** |
|---|---|---|---|---|---|
| $u_1, u_2$ | $sim_{c_1}(u_1,u_2)$ | $sim_{c_2}(u_1,u_2)$ | $sim_{c_3}(u_1,u_2)$ | $sim_{c_4}(u_1,u_2)$ | $sim_{avg}(u_1,u_2)$ |
| $u_1, u_3$ | $sim_{c_1}(u_1,u_3)$ | $sim_{c_2}(u_1,u_3)$ | $sim_{c_3}(u_1,u_3)$ | $sim_{c_4}(u_1,u_3)$ | $sim_{avg}(u_1,u_3)$ |
| $u_1, u_4$ | $sim_{c_1}(u_1,u_4)$ | $sim_{c_2}(u_1,u_4)$ | $sim_{c_3}(u_1,u_4)$ | $sim_{c_4}(u_1,u_4)$ | $sim_{avg}(u_1,u_4)$ |
| $u_1, u_5$ | $sim_{c_1}(u_1,u_5)$ | $sim_{c_2}(u_1,u_5)$ | $sim_{c_3}(u_1,u_5)$ | $sim_{c_4}(u_1,u_5)$ | $sim_{avg}(u_1,u_5)$ |

Choix de k-voisins les plus proches

Prédiction

Une fois toutes les similarités moyennes calculées, le problème du filtrage collaboratif multicritère se réduit au traditionnel filtrage monocritère. Nous choisissons les $k$ voisins dont la similarité avec l'usager cible est la plus grande. Puis nous calculons les prédictions correspondantes avec la formule traditionnelle suivante :

La formule de (2) sera réécrite comme suit :

$$P_{u\,c_i} = \bar{v}_u + \frac{\sum_{j=1}^{k} sim_{avg}(u, u_j)\,(v_{jc_i} - \bar{v}_j)}{\sum_{j=1}^{k} sim_{avg}(u, u_j)} \qquad (4)$$

Avec $k$ : est le nombre de voisins dont la similarité est la plus grande

$k_{voisins}(u)$ : Ensemble des $k$ voisins les plus similaires à $u$ suivant l'approche (HZ).

$u_j \in k_{voisins}(u)$

**Note :** dans ce mémoire nous utiliserons un nombre $k$ = 30 voisins pour calculer la prédiction, comme recommandé dans l'article (Melville, Mooney, & Nagarajan, 2002).

❖ Pseudo-code de l'algorithme HZ:

**Début**

1) Calculer les similarités individuelles entre l'usager cible $u$ et tous les autres usagers $u_j$ pour chaque critère $c_l$ en utilisant la formule (1). Nous obtenons une matrice de similarité $sim_{cl}(u,u_j)$ semblable à Table 4.4
2) Calculer toutes les similarités HZ $Sim_{avg}(u,u_j)$ de l'usager $u$ par rapport aux autres usagers uj en utilisant la formule (3)
3) Choisir $k_{voisins}(u)$, les $k$=30 voisins qui ont la meilleure similarité moyenne $Sim_{avg}(u,u_j)$
4) Calculer la prédiction de chaque critère en utilisant la formule (4)

**Si** usager cible a spécifié un critère $c$

- Classer les articles par ordre décroissant du critère $c$
- Recommander les 10 premiers résultats

**Sinon**

- Classer les articles par ordre décroissant du critère général
- Recommander les 10 premiers résultats

**Fin si**

**Fin**

L'approche HZ construit le voisinage de prédiction en se basant, tout simplement, sur la moyenne des similarités et ne tient compte ni des similarités individuelles relatives au critère à prédire, ni du degré de similarité. En effet, cette approche a deux problèmes. Le premier est que cette approche utilise le même voisinage pour prédire tous les critères. Même si les similarités individuelles, entre l'usager cible et un de ces voisins moyens, sont très divergentes par rapport au critère en cours de prédiction. Cela pourrait fausser les

prédictions pour ce critère. Le deuxième problème est qu'elle ne considère pas un minimum de similarité, car tant qu'elle n'a pas atteint les $k$ voisins, elle continue de prendre des voisins même si ces derniers sont faiblement similaires. Les prochaines approches tentent de résoudre ces deux problèmes.

#### 4.4.4.2 Approche verticale (VL)

L'appellation VL (Verticale) vient de la Table 4.5. Dans cette approche, nous considérons chaque critère comme s'il était unique et nous appliquons le filtrage collaboratif traditionnel.

**Table 4.5** Matrice de similarités et approche VL

| **Paire :** | $c_1$ | $c_2$ | $c_3$ | $c_4$ |
|---|---|---|---|---|
| $u_1, u_2$ | $sim_{c_1}(u_1, u_2)$ | $sim_{c_2}(u_1, u_2)$ | $sim_{c_3}(u_1, u_2)$ | $sim_{c_4}(u_1, u_2)$ |
| $u_1, u_3$ | $sim_{c_1}(u_1, u_3)$ | $sim_{c_2}(u_1, u_3)$ | $sim_{c_3}(u_1, u_3)$ | $sim_{c_4}(u_1, u_3)$ |
| $u_1, u_4$ | $sim_{c_1}(u_1, u_4)$ | $sim_{c_2}(u_1, u_4)$ | $sim_{c_3}(u_1, u_4)$ | $sim_{c_4}(u_1, u_4)$ |
| $u_1, u_5$ | $\boldsymbol{sim_{c_1}(u_1, u_5)}$ | $\boldsymbol{sim_{c_2}(u_1, u_5)}$ | $\boldsymbol{sim_{c_3}(u_1, u_5)}$ | $\boldsymbol{sim_{c_4}(u_1, u_5)}$ |
| | k-voisins | k-voisins | k-voisins | k-voisins |
| | Prédiction | Prédiction | Prédiction | Prédiction |

En d'autres termes, nous considérons un critère à la fois indépendamment des autres. Ainsi, l'item se réduit à un critère, comme dans le cas classique. À tour de rôle, nous déterminons, à l'aide de la formule (1), le voisinage de l'usager cible pour chaque critère à prédire. Donc, pour chaque critère le filtrage collaboratif est monocritère. L'idée derrière cette approche est d'utiliser les $k$ voisins les plus similaires par rapport à ce critère et non pas comme dans la précédente qui utilise les $k$ voisins moyens. Après la détermination du voisinage, nous utilisons la formule (2) pour la prédiction des valeurs inconnues.

❖ Pseudo-code de l'algorithme VL :

**Début**

Calculer les similarités individuelles entre l'usager cible $u$ et tous les autres usagers $u_j$ pour chaque critère $c_l$ en utilisant la formule (1). Nous obtenons une matrice de similarité $sim_{cl}(u,u_j)$ semblable à Table 4.5

**Pour chaque** critère $c_j$

1) Utiliser les similarités individuelles par rapport au critère $c_j$ pour choisir les $k$ = 30 meilleurs voisins
2) Calculer la prédiction du critère $c_j$ en utilisant la formule (2)

**Fin Pour Chaque**

**Si** usager cible a spécifié un critère $c$

- Classer les articles par ordre décroissant des prédictions du critère $c$
- Recommander (afficher) les 10 premiers résultats

**Sinon**

- Classer les articles par ordre décroissant des prédictions du critère global
- Recommander (afficher) les 10 premiers résultats

**Fin si**

**Fin**

Contrairement à l'approche précédente, le voisinage d'un usager cible change en fonction du critère considéré. Cette approche se base sur les similarités individuelles et elle néglige la similarité moyenne. Le fait de choisir les $k$ meilleurs voisins par rapport au critère à prédire, cela ne garantit pas une meilleure prédiction pour ce même critère. Car en réalité l'influence des autres critères n'est pas à écarter. En d'autres termes, un usager avec

une très grande similarité avec l'usager cible perd son poids s'il diverge considérablement par rapport à la majorité des autres critères. Cette fois-ci, la faiblesse vient du fait de négliger la similarité moyenne. C'est un troisième problème qui s'ajoute aux deux premiers préalablement mentionnés dans l'approche précédente HZ. Ajouter à cela, le deuxième problème de l'approche précédente (HZ) qui refait surface dans celle-là. En effet, les deux approches souffrent considérablement quand le voisinage n'est pas très proche de l'usager cible.

Compte tenu des deux approches précédentes, la meilleure solution est de tenir compte des trois problèmes discutés ci-dessus, que nous rappelons sous forme de recommandations comme suit : les nouvelles approches doivent considérer, dans le choix du voisinage de prédiction, à la fois, les similarités individuelles relatives au critère en cours de prédiction, sans négliger la similarité moyenne de ces usagers composant ce voisinage. De plus, nous devrions considérer un *seuil de similarité* pour assurer un minimum de corrélation entre l'usager cible de prédiction et son voisinage, sans pour autant réduire leur nombre. Car plus le seuil est élevé, moins d'usagers similaires passeront cette contrainte. Alors que la taille du voisinage est aussi importante que l'est la similarité par rapport à l'usager cible.

Avant de déterminer le seuil de similarité, d'abord considérons le coefficient de corrélation de Pearson (formule (1)). Une valeur de similarité entre 0.5 et 1 implique une corrélation élevée, et une valeur de similarité entre 0.3 et 0.5, implique une corrélation moyenne. Pour résoudre le problème de voisins faiblement corrélés, nous plaçons un *seuil T* de similarité égal à 0.3, de telle sorte que le voisinage d'un usager cible reste étroit. Néanmoins, l'application d'un seuil unifié aux approches précédentes n'est pas efficace, puisque la taille du voisinage peut ne pas atteindre les *k* voisins, ce qui affecte l'exactitude de la prévision. En effet, la fixation du seuil, pour les approches HZ et VL, a réduit la moyenne de leurs performances. Afin de compléter le voisinage tout en maintenant un haut niveau de similarité, nous proposons les deux prochaines approches.

#### 4.4.4.3 Approche HZ-VL (HoriZontale par la suite VerticaLe)

Il est important de comprendre le rôle du seuil de similarité « *T* » que nous avions introduit ci-dessus. En effet, plus *T* est grand, plus il est difficile de trouver les *k* voisins dont la similarité dépasse ou au moins égale à *T*. Cependant, *T* doit être suffisamment élevé pour avoir une bonne corrélation entre les usagers. Pour cela, nous avons effectué plusieurs essais pour déterminer cette valeur de façon à ne pas sacrifier la qualité de la corrélation.

L'approche HZ-VL (*Horizontal then Vertical*) est une combinaison des approches précédentes (HZ) et (VL) prises dans cet ordre (voir Table 4.6).

**Table 4.6** Matrice de similarités et approche HZ-VL

| **Paire :** | $c_1$ | $c_2$ | $c_3$ | $c_4$ | ***Sim* Moyenne** |
|---|---|---|---|---|---|
| $u_1, u_2$ | $sim_{c_1}(u_1,u_2)$ | $sim_{c_2}(u_1,u_2)$ | $sim_{c_3}(u_1,u_2)$ | $sim_{c_4}(u_1,u_2)$ | $\boldsymbol{sim_{avg}(u_1,u_2)}$ |
| $u_1, u_3$ | $sim_{c_1}(u_1,u_3)$ | $sim_{c_2}(u_1,u_3)$ | $sim_{c_3}(u_1,u_3)$ | $sim_{c_4}(u_1,u_3)$ | $\boldsymbol{sim_{avg}(u_1,u_3)}$ |
| $u_1, u_4$ | $sim_{c_1}(u_1,u_4)$ | $sim_{c_2}(u_1,u_4)$ | $sim_{c_3}(u_1,u_4)$ | $sim_{c_4}(u_1,u_4)$ | $\boldsymbol{sim_{avg}(u_1,u_4)}$ |
| $u_1, u_5$ | $\boldsymbol{sim_{c_1}(u_1,u_5)}$ | $\boldsymbol{sim_{c_2}(u_1,u_5)}$ | $\boldsymbol{sim_{c_3}(u_1,u_5)}$ | $\boldsymbol{sim_{c_4}(u_1,u_5)}$ | $sim_{avg}(u_1,u_5)$ |

Choix de k-voisins les plus proches Avec $\mathbf{Sim_{avg}(u_1,u_j) \geq T}$

Pour chaque critère $c_l$, si le nombre de voisins < k, alors compléter avec les voisins dont $\mathbf{Sim_{cl}(u_1,u_j) \geq T}$

Prédiction

Nous appliquons, premièrement, la méthode HZ sous la contrainte que les *k* voisins doivent dépasser le seuil donné *T* de similarité. Et si le nombre des *k* voisins est inférieur à *k*, nous complétons par les voisins qui ont une similarité individuelle, par rapport au critère en cours de prédiction, supérieure à *T*.

❖ Pseudo-code de l'algorithme HZ-VL :

**Début**

1) Calculer les similarités individuelles entre l'usager cible $u$ et tous les autres usagers $u_j$ pour chaque critère $c_1$ en utilisant la formule (1). Nous obtenons une matrice de similarité $sim_{c1}(u,u_j)$ semblable à Table 4.6
2) Calculer la similarité HZ en utilisant la formule (3)
3) Choisir les $k$=30 voisins dont la similarité moyenne $Sim_{avg}(u,u_j)$est supérieure au seuil de similarité $T$ = 0,3

**Si** le nombre de voisins $N_v$ est inférieur à $k$ = 30
  **Pour chaque** critère $c_1$
    1) Compléter par les $k-N_v$ voisins qui ont une similarité individuelle supérieure au seuil $T$ pour construire l'ensemble des voisins $Vc_j$
    2) Calculer la prédiction du critère $c_1$ en utilisant la formule (2) et le voisinage $Vcj$
  **Fin Pour Chaque**
**Sinon**
  Calculer la prédiction de chaque critère $c_1$ en utilisant la formule (4)
**Fin si**

**Si** usager cible a spécifié un critère $c$
  - Classer les articles par ordre décroissant des prédictions du critère $c$
  - Recommander (afficher) les 10 premiers résultats

**Sinon**
  - Classer les articles par ordre décroissant des prédictions du critère global
  - Recommander (afficher) les 10 premiers résultats

**Fin si**

**Fin**

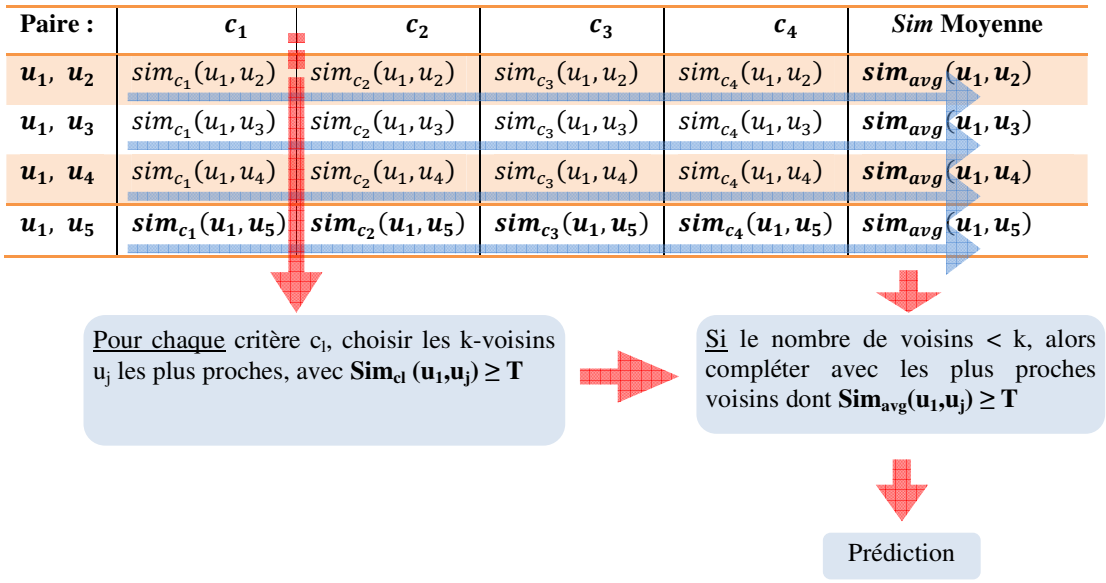
Cette méthode combine en premier les similarités moyennes avec la contrainte $Sim_{avg}(u,u_j) > T$. Si le nombre des voisins de l'usager cible est inférieur à $k$, alors nous complétons par les similarités individuelles correspondant au critère $c_j$ qui est en cours de prédiction avec la contrainte que ces similarités individuelles $Sim_{cj}(u,u_i) > T$. Donc, pour chaque critère $c_j$ à prédire correspond un voisinage que nous obtenons en complétant

l'ensemble de voisins dont $Sim_{avg}(u,u_j) > T$ par les voisins dont la similarité individuelle (par rapport au critère en court $c_j$) vérifie $Sim_{cj}(u,u_i) > T$. Comme dit précédemment, cette approche combine les approches HZ et VL de cet ordre, la combinaison inverse nous donne l'approche suivante.

#### 4.4.4.4 Approche VL-HZ (Verticale par la suite horizontale)

Cette approche ressemble à la précédente, sauf que nous inversons l'ordre d'application des deux autres approches qui entrent dans la combinaison. Donc, nous commençons par calculer les *k* voisins les plus proches à l'usager cible, toujours sous la contrainte que leurs similarités individuelles doivent dépasser le seuil *T*, et si le nombre de voisins est inférieur à *k,* alors nous complétons par les voisins dont la similarité moyenne dépasse ou est égale au seuil *T*.

**Table 4.7** Matrice de similarités et approche VL-HZ



❖ Pseudo-code de l'algorithme VL-HZ :

**Début**

1) Calculer les similarités individuelles entre l'usager cible $u$ et tous les autres usagers $u_j$ pour chaque critère $c_1$ en utilisant la formule (1). Nous obtenons une matrice de similarité $sim_{c1}(u,u_j)$ semblable à Table 4.4
2) Calculer la similarité HZ en utilisant la formule (3)

**Pour chaque** critère $c_1$ à prédire
- Utiliser les similarités individuelles par rapport au critère $c_j$ pour choisir les voisins dont $sim_{c1}(u,u_j) > T$
- **Si** le nombre de voisins $N_v$ est inférieur à $k$
  1) Compléter par les $k-N_v$ voisins qui ont une similarité Horizontale $Sim_{avg}(u,u_j) > T$ pour construire l'ensemble des voisins $Vc_j$
  2) Calculer la prédiction du critère cj en utilisant les formules (2)et(4) ainsi que le voisinage $Vc_j$
- **Sinon**
  - Calculer la prédiction du critère $cj$ en utilisant la formule (2)
- **Fin si**

**Fin Pour Chaque**

**Si** usager cible a spécifié un critère $c$
- o Classer les articles par ordre décroissant des prédictions du critère $c$
- o Recommander (afficher) les 10 premiers résultats

**Sinon**
- o Classer les articles par ordre décroissant des prédictions du critère global
- o Recommander (afficher) les 10 premiers résultats

**Fin si**

**Fin**

Même si les deux dernières approches tiennent compte des précédentes recommandations une autre alternative différente peut résoudre les problèmes, c'est l'objet de cette prochaine approche.

#### 4.4.4.5 Approche HZ-N (*HoriZontal* without *Noise*)

L'approche HZ-N (Horizontale sans Bruit*)* illustrée dans la Table 4.8, est une amélioration ou une optimisation de la première approche horizontale (HZ). Rappelons que cette dernière est basée sur la moyenne des similarités individuelles pour construire le voisinage d'un usager cible $u_i$. Il se trouve que, dans certains cas, malgré la forte corrélation entre cet usager cible $u$ et un certain usager $u_j$ ($Sim_{avg}(u,u_j) \geq T$), les similarités individuelles par rapport au critère $c$ que nous voudrions prédire divergent beaucoup par rapport au seuil $T$. Il est absurde d'utiliser cet usager $u_j$ pour prédire ce critère $c$. Nous qualifions de *bruit* cet usager $c_j$ par rapport à cette approche HZ. En d'autres termes, un voisin est considéré comme un bruit dans la prédiction de la valeur d'un critère $c$, lorsque sa similarité moyenne est haute, et que ses similarités individuelles spécifiques à ce critère $c$ à prédire est faible par rapport au seuil $T$ ($Sim_c(u,u_j) < T$).

**Table 4.8** Matrice de similarités et approche HZ-N

| **Paire :** | $c_1$ | $c_2$ | $c_3$ | $c_4$ | ***Sim* Moyenne** |
|---|---|---|---|---|---|
| $u_1, u_2$ | $sim_{c_1}(u_1,u_2)$ | $sim_{c_2}(u_1,u_2)$ | $sim_{c_3}(u_1,u_2)$ | $sim_{c_4}(u_1,u_2)$ | $\boldsymbol{sim_{avg}(u_1,u_2)}$ |
| $u_1, u_3$ | $sim_{c_1}(u_1,u_3)$ | $sim_{c_2}(u_1,u_3)$ | $sim_{c_3}(u_1,u_3)$ | $sim_{c_4}(u_1,u_3)$ | $\boldsymbol{sim_{avg}(u_1,u_3)}$ |
| $u_1, u_4$ | $sim_{c_1}(u_1,u_4)$ | $sim_{c_2}(u_1,u_4)$ | $sim_{c_3}(u_1,u_4)$ | $sim_{c_4}(u_1,u_4)$ | $\boldsymbol{sim_{avg}(u_1,u_4)}$ |
| $u_1, u_5$ | $\boldsymbol{sim_{c_1}(u_1,u_5)}$ | $\boldsymbol{sim_{c_2}(u_1,u_5)}$ | $\boldsymbol{sim_{c_3}(u_1,u_5)}$ | $\boldsymbol{sim_{c_4}(u_1,u_5)}$ | $sim_{avg}(u_1,u_5)$ |

Choix des k-voisins les plus proches (**$AvgSim(u_1,u_j)$** est la meilleur)

Pour chaque $c_l$ à prédire si **$Sim_{cl}(u_1,u_j) < T$** alors exclure $u_j$ des meilleurs k-voisin de $u_1$

Prédiction

Donc, cette approche se base sur les similarités HZ en éliminant le bruit. Pour chaque critère $c$ à prédire, si la similarité entre l'usager cible et chacun des autres usagers par rapport au critère considéré est faible, nous excluons, tout simplement, l'usager en question du voisinage.

❖ Pseudo-code de l'algorithme HZ-N:

**Début**

1) Calculer les similarités individuelles entre l'usager cible $u$ et tous les autres usagers $u_j$ pour chaque critère $c_l$ en utilisant la formule (1). Nous obtenons une matrice de similarité $sim_{cl}(u,u_j)$ semblable à Table 4.8
2) Calculer toutes les similarités HZ $Sim_{avg}(u,u_j)$ de l'usager u par rapport aux autres usagers $u_j$ en utilisant la formule (3)
3) Choisir $k_{voisins}(u)$, les $k$=30 voisins qui ont la meilleure similarité moyenne $Sim_{avg}(u,u_j)$

**Pour chaque** critère $c_l$ à prédire
- Utiliser les similarités individuelles par rapport au critère $c_l$
- **Pour chaque** usager $u_j$ vérifier
  - **Si** $sim_{cl}(u,u_j) < T$
    - Exclure l'usager $u_j$ du voisinage de prédiction $k_{voisins}(u)$
  - **Fin si**
- **Fin pour chaque**
  - Calculer la prédiction du critère $c_l$ en utilisant la formule(2)et le voisinage $k_{voisins}(u)$ restant

**Fin Pour Chaque**

**Si** usager cible a spécifié un critère $c$
- o Classer les articles par ordre décroissant du critère $c$
- o Recommander les 10 premiers résultats

**Sinon**
- o Classer les articles par ordre décroissant du critère général
- o Recommander les 10 premiers résultats

**Fin si**

**Fin**

Ces trois dernières approches à savoir, HZ-VL, VL-HZ et HZ-N, tiennent comptes des limitations recensées dans les deux premières HZ et VL. À commencer par leur problème commun qui est une absence d'un minimum de similarité entre l'usager cible et les autres usagers. Dans les approches suivantes nous avons comblé ce manque en considérons un seuil de similarité $T$ entre ces usagers. Cette contrainte peut causer un manque considérable de voisins similaires, une chose qui affecte la qualité de la prédiction.

Pour résoudre ce problème et en même temps répondre aux recommandations énoncées après la discussion des deux premières approches, nous avons combiné les similarités individuelles et les similarités moyennes de plusieurs manières avec toujours la contrainte d'une similarité minimale *T*. Pour vérifier ces hypothèses, nous proposons de tester et comparer ces approches dans le prochain chapitre.

## 4.5 Conclusion

Papyres est une application Web de partage, de gestion et de recommandation dédiée spécialement pour les articles de recherche. Elle est au croisement des trois systèmes : systèmes de gestion de références, systèmes de gestion de contenu d'entreprise (ECM) et les systèmes de recommandation d'articles de recherche. Notre conception est basée sur une vision mettant en exergue les différentes parties d'un article de recherche et qui est reflétée sur la majorité de notre système, puisque l'évaluation de ces différentes parties engendre une base de données *multicritère*. Cela nous a amenés ainsi à introduire un nouveau concept qui est celui de la *qualité contextuelle,* qui relativise la notion de la qualité d'un article par rapport au contexte de l'intérêt du chercheur. Une autre conséquence de cette vision et qui constitue une originalité pour ce système est son moteur de recommandation basée sur le *filtrage collaboratif multicritère* qui répond au besoin de cette qualité contextuelle.

L'aspect multicritère d'un article de recherche n'a été exploré que par les types de systèmes de recommandation s'appuyant sur les techniques MCDMs (*Multi-Criteria Decision Making*). À notre connaissance, ce présent travail constitue une première application du filtrage collaboratif multicritère dans le domaine de la recommandation d'articles de recherche. Cette approche traditionnellement monocritère a été étendue pour prendre en charge le multicritère. La considération de cet aspect multicritère est une très récente tendance dans le domaine des systèmes de recommandation, très peu de travaux ont été publiés dans ce sujet. Nous avons contribué à ce domaine avec quatre algorithmes (VL, HZ-VL, VL-HZ et HZ-N) pour améliorer l'approche HZ présentée dans un article pionnier dans ce domaine. Nous proposons de les tester et de les comparer dans le prochain chapitre.

À cette étape, tous les modules composant Papyres ainsi que leur fonctionnement sont présentés. Le prochain chapitre propose deux validations indépendantes. La première consiste en une validation des fonctionnalités proposées comme alternative pour combler le vide dans le domaine entourant l'article de recherche. La deuxième concerne le test des différents algorithmes de filtrage collaboratif multicritère pour comparer leur performance.

# Chapitre 5 Implémentation et validation

Ce chapitre présente les détails de l'implémentation et de la validation de Papyres. Nous commençons par l'environnement de développement et les technologies utilisées. Nous avons inclus des captures d'écran pour illustrer certaines fonctionnalités. Par la suite, nous présentons la procédure suivie pour tester et comparer les différents algorithmes de filtrage collaboratif multicritère ainsi que l'interprétation des résultats obtenus. Pour finir, nous présentons la validation de tout le système grâce au feedback reçu des membres de la communauté de chercheurs participants.

## 5.1 Implémentation de Papyres

Papyres est une application basée sur le Web (Figure 5.1). Elle suit une architecture client-serveur 2-tiers, qui sépare les données, les vues et les contrôleurs.

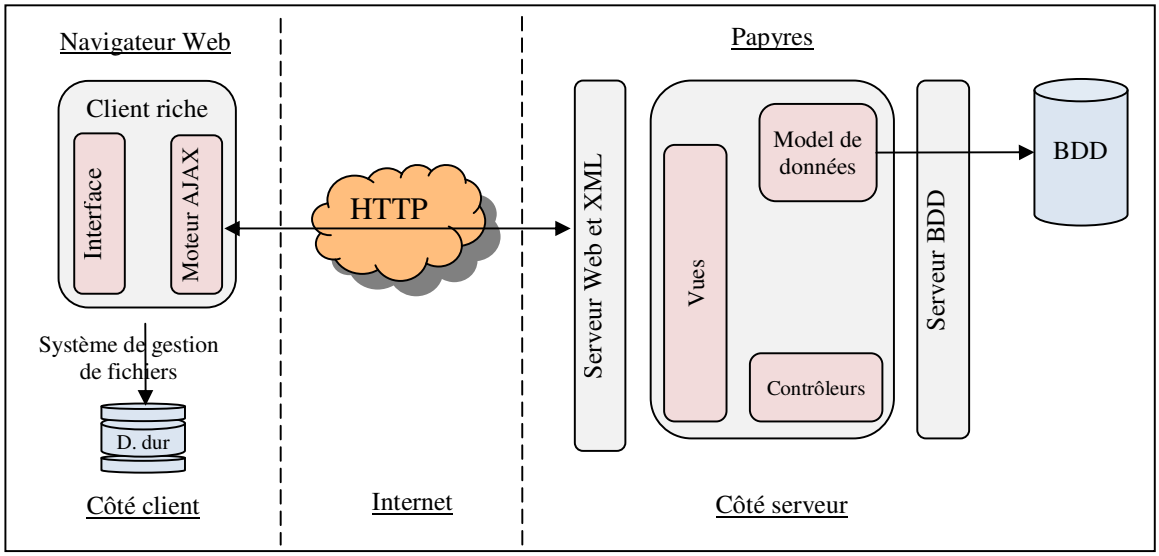



**Figure 5.1** Architecture de l'application

Physiquement, les données sont sur un serveur de données qui est géré par un SGBDR (Système de Gestion de Base de Données Relationnel) dans notre cas MySQL version 3.3.4

et l'application serveur est sur un serveur Web Apache version 5.0.30. L'application côté serveur est dynamique. Elle est implémentée avec le langage de script PHP. Par contre, le côté client est un client riche qui utilise la technologie JavaScript et Ajax pour divers avantages comme, la réduction du trafic, ce qui réalise des économies sur la bande passante, décharge le serveur, et donne un meilleur temps de réponse. Cette technologie nous a permis de reproduire plusieurs fonctionnalités qui sont restées exclusives aux applications locales, comme la gestion de documents, plus particulièrement la gestion de dossiers et sous-dossiers. Nous avons essayé de respecter la séparation entre donnée, traitement et mise en forme en utilisant notamment les recommandations de W3C, comme valider les scriptes XHTML et le CSS.

La prochaine section décrit notre système Papyres, étape par étape. Nous montrerons son fonctionnement global à travers quelques scénarii de fonctionnement et quelques captures d'écrans.

## 5.2 Environnement d'utilisation

La page d'accueil est la page publique de Papyres. Elle contient une brève description de notre application. Cette page offre principalement trois choix : ouverture de session, lien pour rappeler une session oubliée et un lien pour s'enregistrer. Si l'utilisateur possède déjà un compte, il peut directement s'identifier et s'authentifier avec respectivement son nom d'usager et son mot de passe. En cas d'oublie, il peut demander au système de les lui envoyer vers son courriel utilisé lors de son enregistrement la première fois. Sinon, il doit s'enregistrer pour y accéder.

### 5.2.1 Identification et authentification

La première étape pour utiliser le système Papyres, après l'enregistrement, est de s'identifier/authentifier en utilisant le nom d'usager ainsi que son mot de passe. Le système offre un espace privé pour chaque usager ainsi que des services personnalisés, comme la recommandation d'articles de recherche qui est l'un des objectifs de cette application.

### 5.2.2 Ajout d'articles de recherche et disponibilité

Le système offre pour l'usager un espace privé où il peut exploiter le système indépendamment des autres usagers. La première étape dans ce cas est l'ajout d'un article de recherche. L'usager a le choix d'utiliser ses propres ressources ou consulter la base de données publique de Papyres. Cette opération est concrétisée par l'enregistrement des métadonnées de citation de cet article. Idéalement, l'usager pourra attacher une copie de l'article qui sera enregistrée localement et sera accessible par le système. Certaines de ces métadonnées sont obligatoires; elles sont marquées par un astérisque; d'autres sont optionnelles et elles peuvent être ajoutées ultérieurement.

### 5.2.3 Accès et utilisation de l'article

L'usager, après avoir enregistré au moins un article, peut le ou les visualiser sous forme d'une liste ordonnée selon un choix de critères tel que l'ordre alphabétique, l'ordre croissant de l'année de publication.

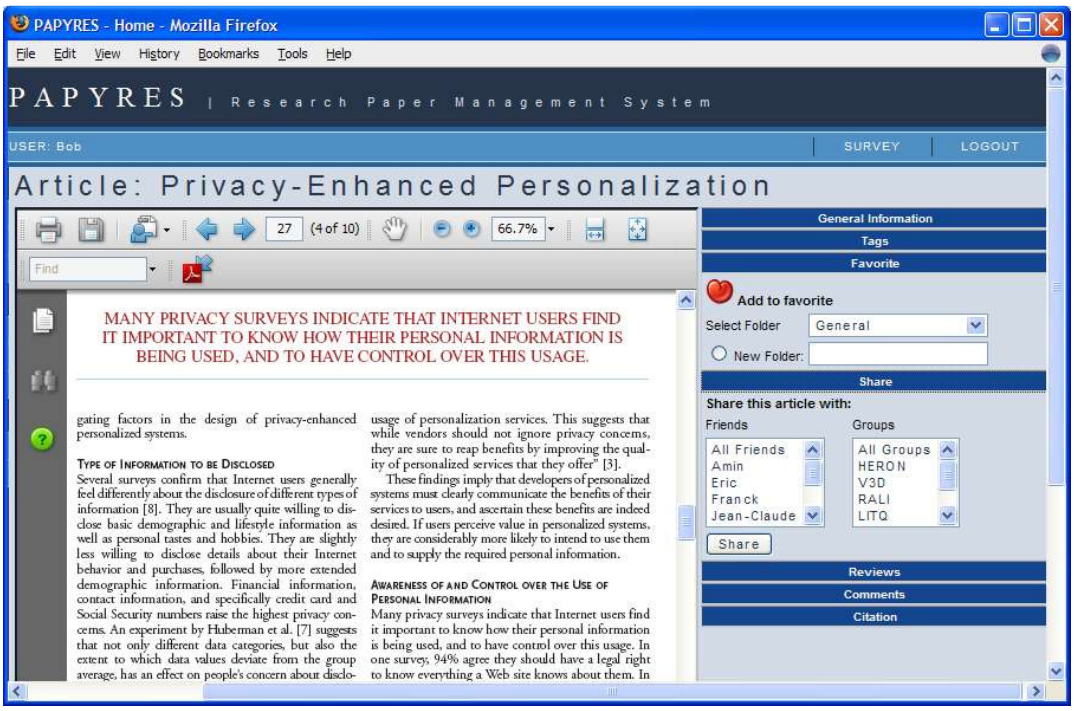



**Figure 5.2** Édition de l'article dans Papyres.

Un article, dont la copie physique est disponible localement, peut être visualisé dans la page Web, dans un endroit réservé à cet effet (voir ci-dessus Figure 5.2). Pendent sa lecture, l'usager peut ajouter différentes informations dites métadonnées Web 2.0, comme des étiquettes (tags), des commentaires des évaluations ou une revue détaillée. Toutes ces notes peuvent être rendues publiques ou être gardées privées (état par défaut). Comme ces notes peuvent être attachées à l'article entier, à l'exception de la revue détaillée, elles peuvent concerner différentes parties de celui-ci, notamment, l'introduction, et l'état de l'art.

La Figure 5.2 représente la page web qui affiche un article préalablement introduit dans Papyres. Le texte de l'article est affiché dans la partie centrale de la page. Les informations relatives à cet article ainsi que les différents outils de prise de notes sont affichés dans la partie droite.

### 5.2.4 Gestion des ressources

Papyres offre une grande souplesse concernant la gestion d'articles (Figure 5.3).

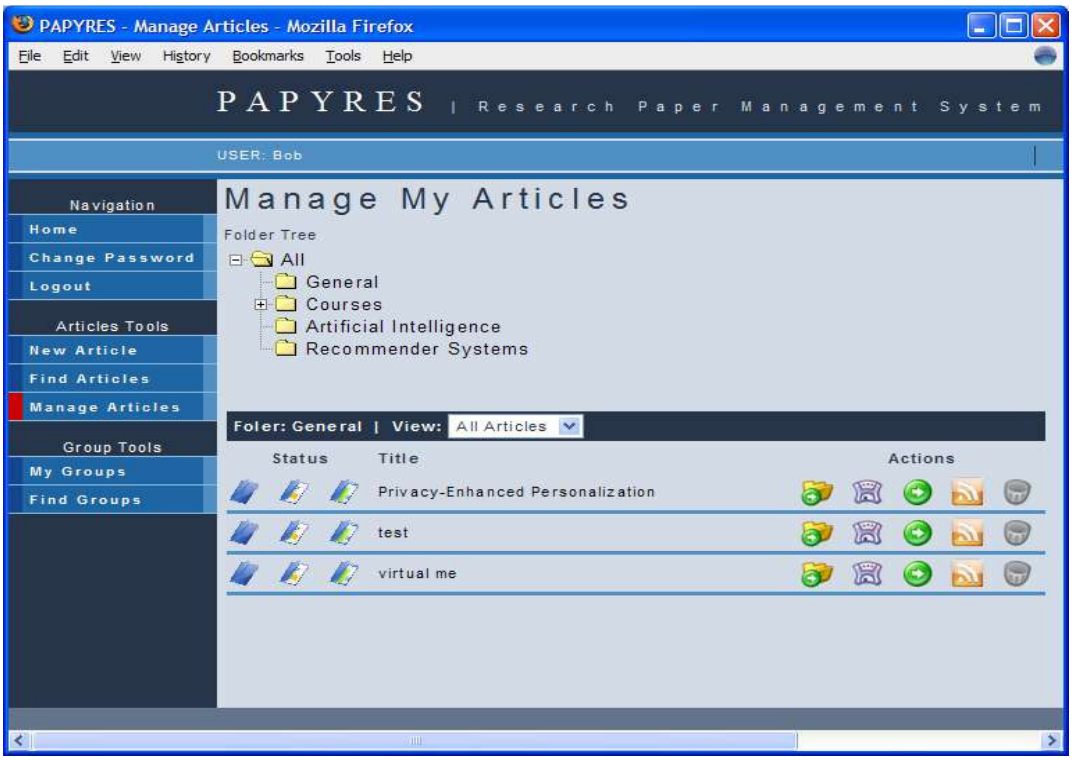



**Figure 5.3** Gestion d'articles de recherche dans Papyres.

En effet, le système permet de classer les différents articles dans des dossiers virtuels. Cela veut dire que l'article proprement dit (son texte) est physiquement dans un emplacement sur le disque, par contre son identifiant est classé dans plusieurs dossiers et sous-dossiers virtuels selon divers critères et dépendamment de la logique de l'usager et de sa volonté. Par exemple, si l'article parle de Système de Recommandation il peut le classer dans un dossier ou sous-dossier correspondant (*Recommender Systems*). Le même article peut être classé dans un sous-dossier du premier ou dans un autre dossier, complètement différent, par exemple : *Artificial Intelligence*.

Le dossier par défaut, où sont rangés tous les articles nouvellement introduits, est appelé « *All* »; c'est la racine de toutes les arborescences crées par les usagers. Avec un simple clic sur un dossier ou sous-dossier, le système affiche son contenu d'articles, dans la partie située au dessous de l'arborescence. Pour chaque article listé, plusieurs informations de suivi sont affichées dans la partie *Status* (à gauche du titre) avec plusieurs actions possibles dans la partie « *Actions* » (à droite du titre). L'usager peut aussi filtrer l'affichage en sélectionnant parmi trois choix prédéfinis : *All Articles*, *Privates*, *Publics* qui affiche respectivement, tous les articles du dossier sélectionné (partagés ou non), seulement les articles privés (non partagés) et seulement les articles publics (partagés).

### 5.2.5 Revue et évaluation d'articles de recherche

Dans Papyres, les chercheurs peuvent faire des revues et évaluer des articles de recherche suivant deux approches complémentaires. La première est l'écriture d'une *revue générale* de l'article dans une zone de revue réservée à cet effet. Le chercheur peut même le faire pour les différentes parties de celui-ci, en utilisant les zones de revue correspondantes. Par exemple, le chercheur peut écrire un texte de revue pour l'*état d'art*, et un texte de revue différent pour l'*approche proposée*. Les sections possibles prédéfinies dans Papyres sont les suivantes : *Introduction*, *État de l'art*, *Méthodologie*, *Expérimentation et validation*, et *les travaux futurs*. Le but d'une telle partition est de fournir une meilleure flexibilité et de faciliter l'évaluation de l'article. En effet, si un chercheur est intéressé spécifiquement par une certaine partie, la consultation des revues qui lui sont dédiées serait assez instructive et bénéfique. La deuxième approche est *le questionnaire d'évaluation*. Cette fois-ci le

chercheur ou l'usager évalue sur une échelle de 1 à 5 les différents aspects de l'article. Parmi les aspects considérés, Nous retrouvons : *Présentation*, *Contribution*, *Orientation technique*, *Niveau Technique*, etc. (voir la section 4.4.2.3). L'utilisation d'une échelle de valeurs numérique dans l'évaluation est bénéfique pour deux raisons. La première est qu'elle fournit pour le chercheur une idée rapide sur les diverses parties de l'article. Deuxièmement, ces valeurs seront exploitées par le système de recommandation (objet de la prochaine section) pour aider des chercheurs à localiser les différentes ressources.

D'autre part, Papyres fournit un *forum de discussion*, où un groupe de chercheurs peut discuter et évaluer des ressources. L'avantage d'une telle approche est double. Tout d'abord, l'utilisation du forum permet aux chercheurs de discuter de façon *asynchrone*, d'évaluer des ressources et d'encourager le partage des connaissances sans la contrainte de se rencontrer ou d'être présents tous en même temps. Deuxièmement, les discussions et les différentes analyses peuvent être facilement stockées et archivées pour de futures références.

### 5.2.6 Localisation de ressources dans Papyres

La recherche d'articles dans Papyres se fait selon différents moyens à savoir : la recherche basée sur les *métadonnées de la citation*, par la spécification d'*étiquettes* (*tags*), par *mots clés* et finalement par le *système de recommandation*. Pour les deux premiers moyens, le système exécute sa recherche dans la base de données. Par contre pour la recherche basée sur les mots clés, le système utilise un filtrage à base de contenu (voir la section 4.4.1). Pour la recommandation, le système effectue d'abord un filtrage à base de contenu, par la suite, il utilise le profil de l'usager pour appliquer le filtrage collaboratif multicritère. Nous avons expérimenté plusieurs approches pour optimiser le choix du voisinage d'un usager cible de recommandation, nous donnerons plus de détails après la comparaison de Papyres avec quelques applications typiques.

## 5.3 Comparaisons

Pour bien montrer les caractéristiques de Papyres, nous l'avons comparé avec les systèmes que nous avions mentionnés dans le chapitre 2 concernant l'état de l'art :

- Sur le plan gestion de citation, nous avons considéré quatre fonctionnalités à savoir : la gestion de citations, le formatage, l'importer/exporter des citations et l'intégrabilité dans *Microsoft Word*.

Comme le montre la Table 5.1, Papyres implémente la majorité des fonctionnalités pour gérer les citations comparativement à EndNote, BibTex tools et CiteULike à la différence qu'il ne s'intègre pas à Microsoft Word. Cette dernière peut être implémentée dans les prochaines versions.

**Table 5.1** Comparaison des outils de citation

| | **Gestion de citation** | **Formatage de citation** | **Import/Export** | **Intégrabilité à Word[1]** |
|---|---|---|---|---|
| **EndNote** | x | x | x | x |
| **BibTeX tools** | x | x | x | x |
| **CiteULike** | x | x | x | |
| **Knowledge Sea II** | | | | |
| **TechLens** | | | | |
| **LiveLink** | | | | x |
| **Documentum** | | | | |
| **Papyres** | x | x | x | |

- Sur le plan de la revue d'articles de recherche, différentes fonctionnalités sont utilisées comme l'affichage de l'*état* de la ressource (lu, commenté, revu.), ajout de commentaires, revue de la ressource et ajout d'étiquettes. Disposition d'un

[1] Microsoft Word

forum afin de discuter les différentes ressources, et de flux RSS pour diffuser toute sorte d'informations et pour établir des listes de surveillances.

La Table 5.2 montre la complétude de Papyres et CiteULike relativement aux autres applications. Néanmoins, il est important de noter que, bien que CiteULike offre l'ajout de commentaires et les fonctions de revue, le niveau de granularité offert par Papyres est plus fin. En effet, CiteUlike offre seulement de commenter globalement les articles de recherche, alors que dans Papyres, les chercheurs peuvent ajouter des commentaires sur diverses parties de l'article.

**Table 5.2** Comparaison de fonctionnalités de revue

| | **État** | **Commentaire** | **Revue** | **Étiquettes** | **Forum** | **RSS** |
|---|---|---|---|---|---|---|
| **EndNote** | | | | | | |
| **Outil BibTeX** | | | | | | |
| **CiteUlike** | x | x | x | x | x | x |
| **Knowledge Sea II** | x | x | x | x | | |
| **TechLens** | x | x | x | x | x | x |
| **LiveLink** | x | x | x | | | |
| **Documentum** | x | x | x | | | |
| **Papyres** | x | x | x | x | x | x |

De plus, dans CiteULike la revue est basée sur un seul critère, qui est *l'évaluation globale* de la ressource, tandis que Papyres offre une *évaluation multicritère* pour évaluer les divers aspects de la ressource.

- Sur le plan organisation de documents, nous avons comparé différentes fonctionnalités fournies pour organiser les articles de recherche, notamment, la possibilité de les insérer dans des *dossiers et sous-dossiers*, prise en charge du

*contrôle d'accès*, la *classification*, la création de *liens personnalisés* pour marquer la relation entre différentes ressources.

La Table 5.3 distingue les applications de gestion de contenu d'entreprise (ECM), en l'occurrence LiveLink et Documentum, avec Papyres qui a bien hérité de leurs fonctionnalités. Mais Papyres se distingue par sa possibilité de gérer des liaisons sémantiques « *Liens Personnalisés* » entre les différents documents lorsqu'un usager juge qu'ils sont liés. Le lien peut être commenté, par exemple l'usager mentionne que tel document est une suite d'un autre, ils sont complémentaires ou ils sont contradictoires.

**Table 5.3** Comparaison de l'organisation des documents

| | **Dossier** | **Sous-Dossier** | **Contrôle d'accès** | **Classification** | **Liens Personnalisés** |
|---|---|---|---|---|---|
| **EndNote** | x | | | | |
| **BibTeX tools** | x | | | | |
| **CiteULike** | | | x | | |
| **Knowledge Sea II** | | | | | |
| **TechLens** | | | x | | |
| **LiveLink** | x | x | x | x | |
| **Documentum** | x | x | x | x | |
| **Papyres** | x | x | x | x | x |

Ces comparaisons montrent la diversité des fonctionnalités de Papyres afin de couvrir un large spectre d'aspects entourant l'article de recherche. Cette richesse est un héritage des trois catégories d'applications étudiées dans l'état de l'art que nous avions rassemblé et complété pour faire un nouveau modèle plus rapproché des besoins de gestion et de recommandation d'articles. Papyres ne se distingue pas des autres applications du genre par seulement ses fonctionnalités, mais par entre autres son système de recommandation. Comme mentionné dans la section 4.4.2.2, le filtrage collaboratif multicritère utilisé dans Papyres est basé sur de nouvelles approches pour choisir le

voisinage de prédiction. Ces approches publiées dans (Naak, Hage, & Aïmeur, 2009) contribuent au domaine des systèmes de recommandation. Avant la validation globale de notre système, nous présenterons d'abord les tests de validation de celles-ci.

## 5.4 Tests des approches de choix du voisinage dans Papyres

Cette section décrit en détail la procédure suivie pour tester les cinq approches de filtrage collaboratif multicritère préalablement introduites dans le chapitre précédent. Tous les tests sont effectués sur un échantillon artificiel suivant l'approche *leave one out approach* (Girouard, Smith, & Slonim, 2006; Mitchell, 1997) qui consiste à choisir aléatoirement une paire Usager/Article puis supposer que cet usager n'a pas encore évalué cet article pour essayer de prédire son évaluation. Ces prédictions seront comparées par la suite avec les valeurs réelles omises en utilisant la mesure MAE (*Mean Absolute Error*) (voir plus loin dans ce chapitre) qui est une mesure de qualité de la prédiction très utilisée dans ce domaine. Toutes les cinq techniques sont testées équitablement sur les mêmes données et nous avons comparé et interprété les résultats obtenus. Avant d'entamer l'analyse et l'interprétation des résultats, nous expliquons, pourquoi nous avons eu recours à se baser sur un échantillon artificiel. Nous détaillons ensuite la méthodologie suivie dans la procédure de test y compris la construction de cet échantillon afin qu'il soit le plus proche de la réalité, tout en étant impartial vis-à-vis des approches concurrentes.

### 5.4.1 Arguments de l'utilisation d'un échantillon artificiel

Le format de l'échantillon, dont nous avons besoin pour tester les différentes approches, doit être multicritère. Car chaque usager évalue chaque article suivant dix critères (voir section 4.4.2.3 ). Nous avons eu beaucoup de difficultés à rassembler un nombre suffisant de participants chercheurs, afin de construire un échantillon de test réel, représentatif et avec la configuration souhaitée. Pour y remédier et comme, les différentes approches sont applicables dans plusieurs domaines, nous avons eu recours à l'Internet, afin de rechercher un échantillon public qui respecte cet aspect multicritère. Nous avons trouvé quelques sites qui utilisent ce genre d'évaluations et nous les avons contactés. Le seul avec qui nos démarches ont abouti est YAHOO! RESEARCH (URL, 13). L'évaluation de films

sur leur site *YAHOO! MOVIES (URL, 14)* est multicritère. Plus précisément, les usagers peuvent noter ou évaluer des films sur leur site Web suivant quatre critères : *Histoire*, *Action*, *Direction* et *Visuels*, ainsi qu'un cinquième critère général qui résume l'ensemble des autres critères.

Les responsables, de YAHOO! RESEARCH, nous ont fait part de leur programme destiné à la communauté de chercheurs et de leur volonté à nous aider en nous envoyant leur catalogue[1] d'échantillons disponibles, sous réserve de satisfaire certaines conditions. Après maints échanges de courriels, ils ont fini par nous envoyer un échantillon que nous avions choisi préalablement du catalogue. À notre surprise l'échantillon n'était pas formaté suivant nos besoins, toutes les valeurs des critères ont été omises et il ne restait que le critère général. Après les avoir recontactés, ils nous ont proposé d'attendre la publication du nouveau catalogue[2] contenant plusieurs nouveaux échantillons. Effectivement, ils nous l'ont envoyé et une fois de plus nous avons constaté qu'aucun des nouveaux éléments ne répond à nos besoins. Nous les avons recontactés pour nous en assurer et voici la réponse intégrale de leur responsable : « *I just chatted with the research scientist and as I was afraid, we don't have the data in the configuration that you'd like* ». D'où notre recours à la construction d'un échantillon artificiel. D'après les auteurs de l'article de journal (Herlocker, Konstan, Terveen, & Riedl, 2004), qui est une grande référence dans le domaine des systèmes de recommandation, l'utilisation d'un échantillon artificiel peut servir pour des résultats préliminaires en attendant d'autres essais basés sur des données réelles.

### 5.4.2 L'échantillon de test (*Dataset*)

L'échantillon, sur lequel sont effectués les tests, a été construit artificiellement d'une manière pseudo-aléatoire[3], suivant deux étapes. Premièrement, nous avons créé un ensemble de 20 usagers différents. Pour chaque usager, nous avons spécifié aléatoirement 30 évaluations différentes. Le caractère aléatoire de ces valeurs induit une absence de

---

[1] Catalogue : « *Yahoo! Research Webscope Data Sets*»
[2] Catalogue : « *Yahoo! Webscope Datasets Catalog January 2009*»
[3] Afin de diminuer l'effet aléatoire, nous avons tenu compte de liaisons logiques entre les évaluations.

corrélation logique entre les évaluations des usagers pour ces articles, et c'est pourquoi cette façon de faire ne reflète pas la réalité de la vie courante et ne répond pas aux besoins de notre expérience. Par exemple, considérons le cas d'un usager *u* qui évalue deux critères $c_1$ et $c_2$ d'un même article et supposons que ces deux critères sont liés d'une certaine liaison logique. Le fait de leur donner aléatoirement les valeurs 1 et 5 introduit une incohérence et cette liaison n'aura plus son effet sur la prédiction. Ou encore, si nous considérons le cas de deux usagers qui ont donné une même évaluation globale pour un certain article, mais une évaluation opposée pour un certain critère, qui logiquement devrait être similaire, par conséquent, cela affectera les résultats du test. Pour diminuer cet effet aléatoire des évaluations, nous avons augmenté l'échantillon de départ de sorte à simuler cette liaison logique entre les évaluations des usagers, de la manière suivante : pour chacun des 20 usagers de départ, nous avons créé 10 nouveaux autres pour lesquels leur évaluations sont basées sur le premier, et sont variées d'une façon logique, mais aléatoire. Précisément, considérons $R_{a,i}$ l'ensemble des évaluations de l'usager initial *a* pour un article *i*. Pour obtenir l'ensemble des évaluations $R_{b,i}$ du nouvel usager *b*, nous nous basons sur l'usager initial *a* de la manière suivante : $r_{b.i} = r_{a.i} + x$ avec *x* est un entier pris aléatoirement de l'ensemble X = {-1,0,1}, qui est aussi pris aléatoirement parmi les ensembles suivants : {-1,0,1}, {-2,-1,0}, {0,1,2}, {-3,-2,-1}, {1,2,3}, {-1,0,1,2}, {-2, -1,0,1}. Nous nous assurons que chaque ensemble est choisi au moins une fois. De cette façon, nous obtiendrons pour chacun des 20 usagers initiaux 10 autres (*10 x 200* autres usagers) et cela nous donne, au total, un échantillon de 220 usagers différents qui chacun d'eux a évalué 30 articles différents.

### 5.4.3 MAE (*Mean Absolute Error*)

Pour comparer les cinq approches : HZ, VL, HZ-VL, VL-HZ, et HZ-N, nous avons procédé comme suit : nous avons choisi aléatoirement un chercheur ou un usager ainsi qu'un article évalué par cet usager. Nous faisons abstraction de cette évaluation en supposons qu'il ne l'a pas encore évalué et nous essayons de prédire ses évaluations. À la fin nous comparerons les valeurs des deux évaluations : celles originales avec celles prédites et cela en utilisant la mesure MAE (*Mean Absolute Error*) (formule (5)).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \quad (5)$$

Où : n est le nombre de prédictions,

$f_i$ est la prédiction $i$ et

$y_i$ est l'évaluation originale.

MAE est une métrique qui est régulièrement utilisée pour évaluer la précision d'une telle prédiction. En bref, elle représente la différence moyenne entre la valeur originale et celle prédite.

### 5.4.4 Test du système de recommandation et résultats

Pour tester les approches, nous avons sélectionné aléatoirement un ensemble de test composé de 100 paires chercheur/article. Par la suite, les cinq approches sont utilisées pour prédire les évaluations de l'ensemble de test. Ces dernières sont comparées avec les évaluations originales en utilisant la méthode MAE. Le MAE de chaque critère est enregistré, ainsi que la moyenne des MAE à traves tous les critères. La moyenne MAE à travers les 100 itérations est utilisée pour comparer la performance des différentes approches implémentées. La Figure 5.4 montre le meilleur cas, le pire et le cas moyen respectivement sur les 100 itérations des cinq approches : minimum MAE (MIN MAE), maximum MAE (MAX MAE) et la moyenne MAE (AVG MAE).

En général, l'approche VL est la moins performante, elle souffre principalement lorsque la similarité entre un usager et son voisinage n'est pas suffisamment rapprochée. Ce problème est abordé par l'approche HZ qui considère la similarité globale. Dans ce cas, même si le voisinage d'un usager est un peu loin par rapport à un certain critère, la moyenne des similarités réduit l'erreur moyenne dans la prédiction. D'un côté, les approches HZ-VL et VL-HZ maximisent la similarité du voisinage, et offrent une meilleure performance générale par rapport à HZ et VL utilisés séparément. De l'autre côté, malgré que HZ-N a le plus haut MAX MAE, cette approche reste celle qui offre la meilleure

performance générale. HZ-N prend son avantage de la similarité générale avec la réduction du *bruit* induit par le voisin dont la similarité générale est grande, mais qui n'est pas très similaire pour un certain critère. Dans le but d'avoir une meilleure interprétation des valeurs MAE, il est important de considérer l'échelle d'évaluation utilisée dans l'évaluation des articles.

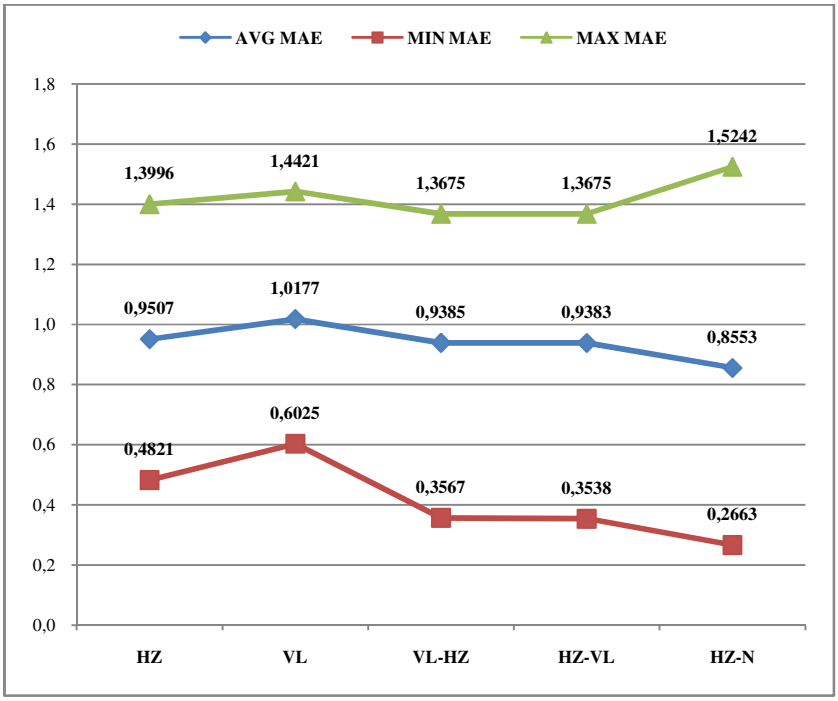



**Figure 5.4** Comparaison de MAE

En effet, une MAE de 0.5 indique que les prédictions, diffèrent de 0.5 de l'évaluation originale. Afin d'évaluer l'impact de cette différence, il est important de considérer l'échelle d'évaluation utilisée, dans la prédiction. En effet, une différence de 0.5 sur une échelle de 1 à 5 est plus significative qu'une échelle de 1 à 20. En d'autres termes, une différence de MAE de 0.5 sur une échelle de 1 à 5 représente une variation de 10 % alors que cette différence sur une échelle qui s'étale de 1 à 20 ne représente qu'une variation de 2.5 % et, par conséquent, un plus faible impact sur la précision.

La Figure 5.5 (ci-dessous) montre une interprétation de la moyenne MAE pour chacune des cinq approches. Les variations MAE sont représentées en pourcentage pour

montrer visuellement son impact par rapport à l'échelle d'évaluation de 1 à 5. Bien que les résultats soient encourageants et que l'approche HZ-N offre une amélioration de la précision sur les autres approches, une valeur MAE de 0,8 (qui représente une variation de 17 %) n'est pas complètement satisfaisante. Néanmoins, nous pensons que le MAE de 0,8 est dû en grande partie au fait que les données sont générées de façon aléatoire (ou pseudo aléatoire).

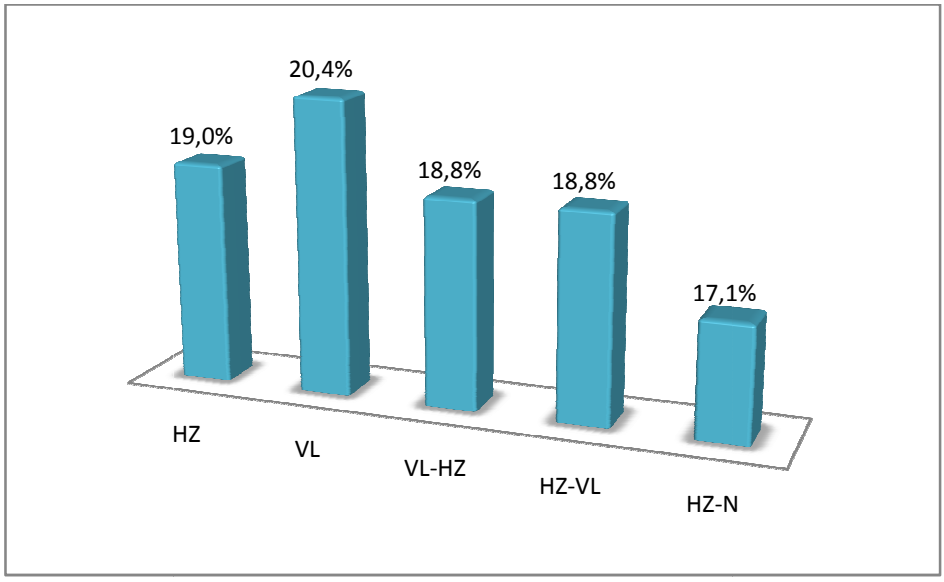



**Figure 5.5** Interprétation des MAE moyenne

En effet, nous pensons que la valeur du MAE sera moindre lors de l'essai des approches sur une véritable base de données. En outre, nous pensons également que l'approche Horizontale HZ-N sera plus efficace que les autres approches, puisque tous les tests ont été exécutés sur le même ensemble de données.

### 5.4.5 Récapitulatif

Bien que l'utilisation de données générées pour tester les performances des différentes approches ne soit pas la méthodologie la plus souhaitable, néanmoins nous argumentons cette procédure pour valider des résultats préliminaires. En effet, en comparant les résultats de plusieurs approches sur un même échantillon de données, même

si celui-ci est généré d'une façon pseudo-aléatoire, il offre une certaine validité à ces résultats. De plus, le caractère aléatoire donne une neutralité à l'échantillon, puisqu'il n'est pas conçu spécifiquement pour améliorer la performance d'une approche, pendant qu'il détériore la performance d'une autre.

En outre, l'aspect pseudo-aléatoire de l'échantillon assure une bonne répartition des profils et des similitudes sur un large éventail. En particulier, il est fondé sur la corrélation de Pearson (équation (1)) qui varie entre -1 (complètement différent) et 1 (parfaitement similaire), tandis que la similarité entre les utilisateurs pseudo-aléatoires varie entre 0,941 et -0,804, couvrant ainsi la plus grande partie du spectre de valeurs.

Toutefois, nous sommes d'accord avec les auteurs de l'article (Herlocker, Konstan, Terveen, & Riedl, 2004), que les données générées peuvent être utilisées pour des résultats *préliminaires* et que d'autres essais devront être effectués pour des résultats conclusifs et concluants.

## 5.5 Validation globale de Papyres

Le processus de validation a été divisé en deux étapes. Dans la première étape, nous avons demandé aux chercheurs participants de remplir un questionnaire qui concerne leurs habitudes de recherche. Nous leur avons proposé quelques questions, incluant, comment ils organisent leurs ressources, est-ce qu'ils utilisent un système de gestion bibliographique. Après avoir complété le questionnaire, les répondants ont eu l'opportunité de voir et d'utiliser les différentes fonctionnalités de Papyres. Dans la deuxième étape en terminant, les répondants donnent une évaluation générale pour le système par rapport aux fonctionnalités fournies, et la facilité de son utilisation. De plus, ils ont eu la possibilité d'exprimer des commentaires et/ou des suggestions concernant des fonctionnalités manquantes qui pourraient y être intégrées. Un total de 83 répondants a participé dans le processus d'évaluation : 13 professeurs, 27 étudiants au doctorat, et 43 étudiants en maîtrise. Un effort particulier est fourni pour présenter à chacun des participants les objectifs et le but de cette application.

Les résultats de la validation renforcent l'hypothèse sur laquelle est basée Papyres : les chercheurs ont besoin d'un système de gestion d'articles de recherche qui les aidera à tirer plus d'avantages de leurs ressources en recherche. En effet, comme le montre la Figure 5.6, la plupart des chercheurs interrogés organisent directement leurs articles dans des dossiers rigides.

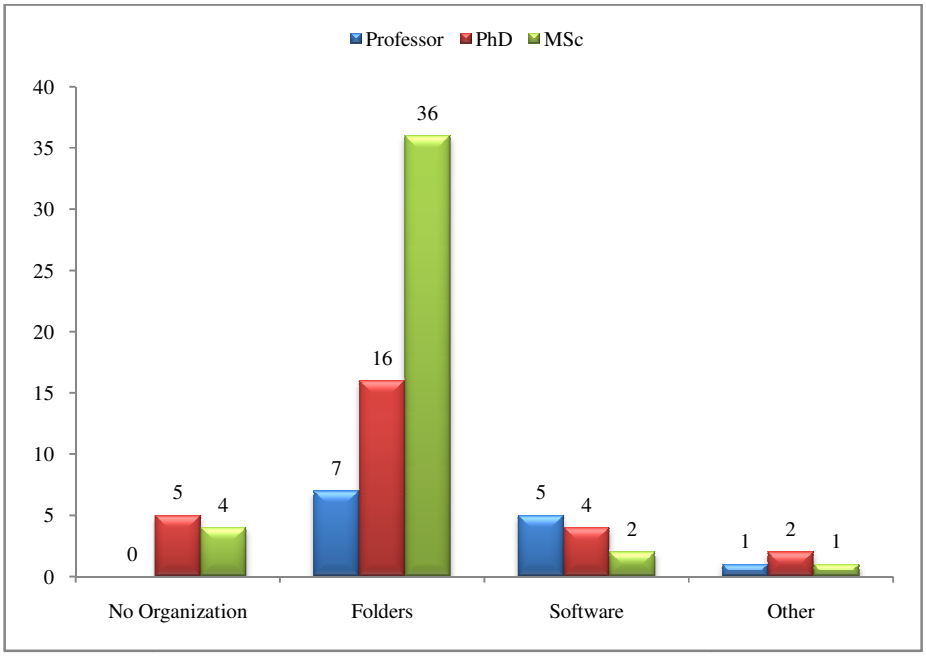



**Figure 5.6** Habitudes d'organisation

Les répondants, qui ont déclaré avoir utilisé un logiciel pour organiser leurs ressources, se sont servis pour ça de systèmes de gestion de bibliographie. Ces systèmes ont leurs avantages par rapport à l'organisation manuelle des ressources dans le dossier, mais restent encore limités.

En outre, lorsque nous leur avons demandé d'évaluer sur une échelle de 1 à 5 (1 : Ne jamais, 5 : Toujours), la question nº 5 : « *Prenez-vous des notes sur les articles que vous lisez ?* », la plupart ont répondu qu'ils le font régulièrement. Pourtant quand nous leur avons demandé d'évaluer sur la même échelle la question nº 6 : « *Utilisez-vous un système pour gérer la prise de note sur ces articles ?* C'est-à-dire, s'ils utilisent un logiciel pour gérer leurs prises de notes sur les articles, la plupart ont répondu jamais (Figure 5.7).

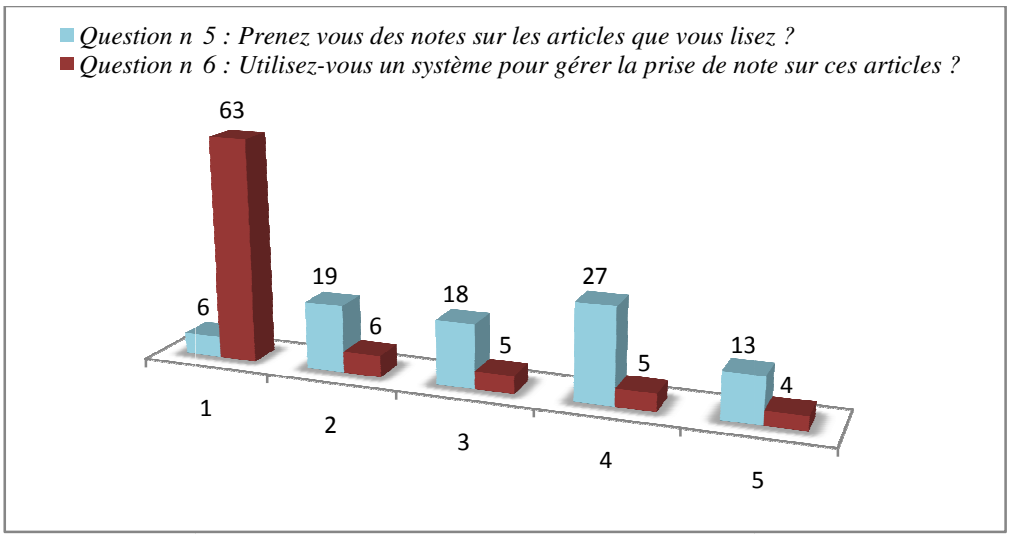



**Figure 5.7** Prise de note et organisation

Quand nous avons demandé d'évaluer sur une échelle de 1 à 5 (1 : jamais, 5 : toujours), combien de fois, quand vous cherchez de nouveaux articles, vous vous êtes intéressés à une partie d'un article, comme l'état de l'art. La majorité des répondants disent l'avoir été de nombreuse fois et d'une façon régulière (Figure 5.8).

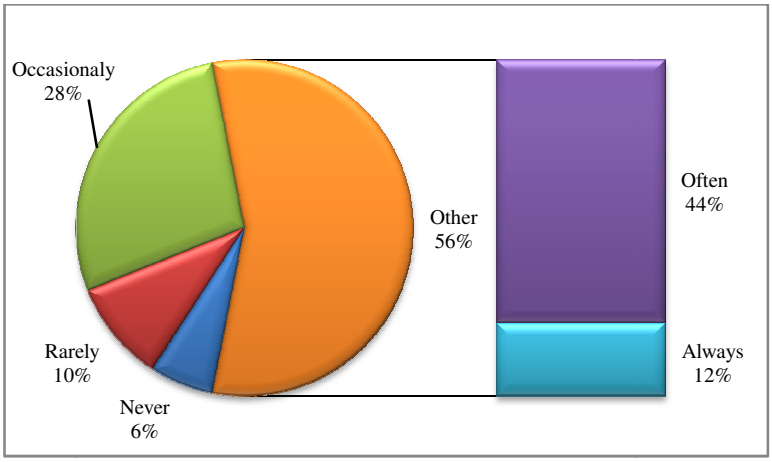



**Figure 5.8** Intérêt de recherche dans une partie d'un article

Finalement après la le test de Papyres, nous leur avons demandé d'évaluer les différentes fonctionnalités qu'il offre. La moyenne des évaluations récoltées est proche de 5, exactement 4,43 sur une échelle 1 à 5 (1 : Très pauvre, 5 : Excellent). En outre, lorsque

nous leur avons demandé d'évaluer sur une échelle de 1 à 5 (1 : Jamais, 5 : Toujours) si Papyres les encourage à : (a) gérer les ressources, (b) partager des ressources, et (c) partager les observations et commentaires, les réponses, en moyenne, sont respectivement : 4.52, 4.42 et 4.51. Enfin, lorsque nous leur avons demandé quelles autres fonctionnalités ils aimeraient ajouter à Papyres, une suggestion intéressante qui revient est d'inclure des *représentations visuelles des relations ressource/auteur*. Plus précisément, ces suggestions proposent de représenter visuellement le plus de sujets communs par un certain auteur (semblable à un nuage d'étiquettes), ou un graphique mettant en évidence les citations entre les articles, c'est-à-dire en utilisant un graphe orienté pour représenter les relations entre les articles, où les sommets représentent les articles et les arêtes correspondent aux citations.

## 5.6 Conclusion

Les résultats de test de Papyres sont très encourageants, que ce soit sur le plan fonctionnalités et environnement de travail, ou sur le plan algorithmes de recommandation. Sur le plan fonctionnalité, 83 répondants ont testé et évalué Papyres. La moyenne de toutes les évaluations, concernant la question relative aux fonctionnalités offertes par notre système, était de 4.43 sur l'échelle allant de 1 : Très pauvre à 5 : Excellent. Sur le plan de la recommandation, les approches que nous avions proposées ont donné de meilleurs résultats par rapport à l'approche horizontale (HZ). Particulièrement, l'approche horizontale HZ-N (*HoriZontal without Noise*) est celle qui a le mieux fonctionné. Certes, ces résultats ne sont pas définitifs à cause de l'échantillon de données artificiel, mais nous avons argumenté notre procédure en nous appuyant sur l'étude proposée dans l'article (Herlocker, Konstan, Terveen, & Riedl, 2004) et tous les tests ont été menés sur un même échantillon qui est constitué aléatoirement de façon à ne pas favoriser une approche particulière au détriment d'une autre. Plus encore, nous avons essayé de minimiser l'effet aléatoire afin de mieux simuler des données réelles. Compte tenu de ces résultats, nous avons défini un nouveau cadre pour la gestion d'articles de recherche avec d'excellentes perspectives futures.

# Chapitre 6 Conclusion

La problématique abordée dans ce mémoire est centrée au tour de l'*article de recherche* avec toutes les informations qui lui sont liées. Précisément, dans un contexte de recherche scientifique, les chercheurs font face à de nombreuses contraintes qui sont liées à la gestion et à la localisation de ces ressources. Les solutions existantes, notamment : les systèmes de gestion de références, les systèmes de gestion de contenu d'entreprise, et les systèmes de localisation d'articles de recherche ne répondent pas aux besoins particuliers de notre objet de recherche, mais chacune présente des fonctionnalités intéressantes à implémenter et à adapter pour les spécificités de l'article de recherche. Dans une première étape, nous avons pensé à regrouper ces fonctionnalités dans un cadre unique centré autour de cet article et nous avons appelé cette nouvelle classe de systèmes : système de gestion et de recommandation d'articles de recherche. Pour illustrer cette classe, nous avons implémenté Papyres (Naak, Hage, & Aïmeur, 2008, 2009), une application web de partage de connaissances liées aux articles de recherche, tout en implémentant un ensemble de fonctionnalités qui s'apparentent aux trois types de systèmes précédents.

Dans une seconde étape, Papyres exploite une vue originale de l'article de recherche, c'est la *vue par parties*. Spécifiquement, l'article de recherche est considéré par Papyres comme une entité divisible qui est repartie suivant un nombre de parties prédéfinies : Introduction, État de l'art, Méthodologie, Expérimentation et validation, et Travaux futurs. Cette nouvelle vision introduite dans Papyres permet une flexibilité tant recherchée et constitue une réponse au sondage que nous avions mené [Naak, Hage *et al.,* 2008] où la majorité des répondants disent avoir toujours été intéressés par une partie spécifique d'un article, lors de leurs opérations de recherche. Ainsi, il est possible de commenter, étiqueter et évaluer séparément des parties prédéfinies, et non pas la totalité, de l'article. Dès lors, l'évaluation de l'article de recherche dans notre système se fait selon

plusieurs critères, parmi lesquels ceux attribués à ces parties prédéfinies, mettant ainsi en évidence l'*évaluation multicritère,* à travers laquelle un chercheur peut exprimer ses préférences de qualité à un niveau de granularité plus fin, agissant à l'échelle des parties de l'article. Par conséquent, la qualité d'un article n'est pas reflétée uniquement par son évaluation globale, mais elle est aussi liée aux qualités intrinsèques de ses parties.

Cette façon de voir a mené à introduire un nouveau concept, celui de la *qualité contextuelle* qui reflète la qualité d'un article relative à un contexte de recherche donné et qui est exprimée explicitement par le chercheur en indiquant la partie de l'article qui l'intéresse. Plus encore, nous avons extrapolé ce concept vers le système de recommandation pour prendre en charge ce nouveau facteur avec l'introduction de la *recommandation multicritère*. Plus précis, cette dernière est concrétisée par sa partie qui est le *filtrage collaboratif multicritère*. Au meilleur de notre connaissance, cette dernière approche n'a jamais été appliquée dans le domaine de la recommandation d'articles de recherche, ce qui constitue une autre originalité pour nos travaux. La recommandation multicritère indépendamment du domaine d'application a été introduite d'une façon générale dans l'article (Adomavicius & Tuzhilin, 2005) en tant que possible extension du domaine de la recommandation, par la suite l'article (Adomavicius & Kwon, 2007) précise concrètement des algorithmes pour mettre en œuvre cette nouvelle tendance. En bref, leurs travaux consistaient en l'extension des algorithmes traditionnellement monocritères pour le cas multicritère. Nous nous sommes intéressés particulièrement au filtrage collaboratif et nous avons proposé quatre nouvelles approches pour optimiser le choix du voisinage : VL(*VerticaL*), HZ-VL(*HoriZontal* then *VerticaL*), VL-HZ(VerticaL then *HoriZontal*), HZ-N(*HoriZontal* without *Noise*). Ces approches constituent une autre contribution au domaine de la recommandation multicritère (Naak, Hage, & Aïmeur, 2009). Ces quatre approches, augmentées de HZ (*HoriZontal* ou agrégation de similarités) (Adomavicius & Kwon, 2007), ont été testées et comparées. Les résultats ont été significatifs et nous avons noté la performance de nos approches. En particulier l'approche HZ-N est celle qui a donné les meilleurs résultats.

Nous sommes conscients que ces résultats ne sont pas définitifs, mais préliminaires à cause de l'aspect artificiel de l'échantillon de données. Sachant la difficulté qui consiste à

réunir un échantillon réel et représentatif qui répond à nos besoins expérimentaux et la difficulté de les trouver dans notre contexte de recherche qui est l'article de recherche, nous avons pensé à d'autres domaines d'application, puisque ces algorithmes sont polyvalents et à condition que l'évaluation soit multicritère. Après une recherche sur le web, nous avons eu recours à YAHOO! RESEARCH qui met à la disposition de la recherche académique plusieurs échantillons réels pour des besoins d'expérimentations. Malheureusement, aucun de ces derniers n'est multicritère. En dernier recours, nous nous sommes appuyés sur l'article (Herlocker, Konstan, Terveen, & Riedl, 2004) pour argumenter notre choix et travailler sur des données que nous avions générées. Comme suggérés par cette étude, des tests futurs devraient être effectués sur des données réelles avant de tirer des conclusions finales. Néanmoins, la façon dont est généré l'échantillon offre une certaine crédibilité pour les résultats obtenus, puisque toutes les approches ont été testées sur un même échantillon généré pseudo aléatoirement de telle sorte à ne pas favoriser une approche au détriment d'une autre. Comme dit précédemment, cette nature pseudo aléatoire atténue l'effet de la dispersion des données et leur offre une certaine corrélation logique. C'est ce que démontre la variation de la similarité dans notre exemple qui varie entre -0,804 et 0,941, couvrant ainsi la majorité du spectre de variations de la corrélation de Pearson (équation (1)).

En général, Papyres à obtenu des résultats très encourageants lors de sa soumission pour des tests auprès d'une communauté de chercheurs composée de 13 professeurs, 27 étudiants Ph.D, et 43 étudiants en M.Sc. En effet, lorsque nous leur avons demandé d'évaluer les fonctionnalités offertes par ce système, la moyenne des évaluations était 4,43/5. Et comme l'espace des fonctionnalités est vaste, nous espérons avoir défini un cadre de travail novateur ouvrant ainsi plusieurs possibilités pour améliorer ce genre de systèmes.

En guise de travaux futurs, nous commencerons par quelques suggestions reçues pendent le processus de validation pour améliorer notre système sur le plan fonctionnalités, parmi celles-ci la possibilité de visualiser graphiquement des relations auteur/ressource. Sur le plan recommandation, afin de compléter le test des approches de recommandation, il est important de les tester sur un échantillon de données réelles pour aboutir à des conclusions finales. D'autres extensions sont intéressantes pour améliorer la précision de la recommandation notamment l'introduction d'autres dimensions dans la recommandation

avec l'enrichissement du profil de l'usager. Par exemple, inclure le niveau de confiance par rapport au domaine de recherche d'un article, et l'expérience dans le processus de revue, qui peuvent respectivement départager l'ensemble des données selon les dimensions domaine d'expertise et niveau d'expérience.

# Bibliographie


Adomavicius, G., & Kwon, Y. (2007). New Recommendation Techniques for Multicriteria Rating Systems. *IEEE Intelligent Systems,* 22(3), 48-55.

Adomavicius, G., Sankaranarayanan, R., Sen, S., & Tuzhilin, A. (2005). Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Trans. Inf. Syst.,* 23(1), 103-145.

Adomavicius, G., & Tuzhilin, A. (2001). Multidimensional Recommender Systems: A Data Warehousing Approach, *Proceedings of the Second International Workshop on Electronic Commerce*: Springer-Verlag.

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering,* 17(6), 734-749.

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

Balabanovi, M., & Shoham, Y. (1997). Fab: content-based, collaborative recommendation. *Commun. ACM,* 40(3), 66-72.

Basu, C., Hirsh, H., Cohen, W., & Manning, N. (2001). Technical paper recommendation: A study in combining multiple information sources.

Belkin, N. J., & Croft, W. B. (1992). Information filtering and information retrieval: two sides of the same coin? *Commun. ACM,* 35(12), 29-38.

Bollacker, K., Lawrence, S., & Giles, L. (1998). *CiteSeer: An Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications.* Communication présentée Proceedings of the Second International Conference on Autonomous Agents, Minneapolis, Minnesota, United States.

Bollacker, K., Lawrence, S., & Giles, L. (1999). A System for Automatic Personalized Tracking of Scientific Literature on the Web. Dans DL '99 (Éd.), *The Fourth ACM Conference on Digital Libraries* (pp. 105-113). Berkeley, California, United States: ACM.

Breese, J. S., Heckerman, D., & Kadie, C. (1998). Empirical Analysis of Predictive Algorithms for Collaborative Filtering, *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence* (pp. 43-52).

Brusilovsky, P., Chavan, G., & Farzan, R. (2004). Social Adaptive Navigation Support for Open Corpus Electronic Textbooks. *Adaptive Hypermedia and Adaptive Web-Based Systems* (pp. 24-33).

Brusilovsky, P., Farzan, R., & Jae-wook, A. (2005). Comprehensive personalized information access in an educational digital library, *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 05)* (pp. 9-18). Denver, USA.

Burke, R. (2000). Knowledge-Based Recommender Systems. Dans Kent, A. (Éd.), *Encyclopedia of Library and Information Systems* (Vol. 69): Marcel Dekker.

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

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

Girouard, A., Smith, N. W., & Slonim, D. K. (2006). *Motif Evaluation by Leave-one-out Scoring.* Communication présentée Computational Intelligence and Bioinformatics and Computational Biology, 2006. CIBCB '06. 2006 IEEE Symposium on.

Goldberg, D., Nichols, D., Oki, M. B., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Commun. ACM,* 35(12), 61-70.

Herlocker, J., L., Konstan, J., A., Terveen, L., G., & Riedl, J., T. . (2004). Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.,* 22(1), 5-53.

IEEE Learning Technology Standards Committee. (2002). *Standard for Learning Object Metadata*.

Kampffmeyer, U. (2006). *ECM Enterprise Content Management*. Hamburg.

Kapoor, N., Chen, J., Butler, J., T., Fouty, G., C., Stemper, J., A., Riedl, J.*, et al.* (2007). Techlens: a researcher's desktop. Dans '07, R. (Éd.), *2007 ACM conference on Recommender systems* (pp. 183-184). Minneapolis, MN, USA: ACM.

Manouselis, N., & Costopoulou, C. (2007). Analysis and Classification of Multi-Criteria Recommender Systems. *World Wide Web,* 10(4), 415-441.

Mao, Y., Vassileva, J., & Grassmann, W. (2007). *A System Dynamics Approach to Study Virtual Communities.* Communication présentée System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on.

MARKESS, I. (2008). *Gestion de Contenu d'Entreprise : Enjeux & Perspectives*. Paris, France.

Matsatsinis, N. F., Lakiotaki, K., & Delias, P. (2007). A System based on Multiple Criteria Analysis for Scientific Paper Recommendation. Dans '07, P. (Éd.), *11th Panhellenic Conference in Informatics* (pp. 135-149). Patras, Greece.

Melville, P., Mooney, R. J., & Nagarajan, R. (2002). Content-boosted collaborative filtering for Improved recommendations. Dans Dechter, R., Kearns, M. & Sutton, R., Eds (Éds.), *Eighteenth National Conference on Artificial Intelligence* (pp. 187-192). Edmonton, Alberta, Canada: American Association for Artificial Intelligence.

Middleton, S. E., Shadbolt, N. R., & De Roure, D. C. (2004). Ontological user profiling in recommender systems. *ACM Trans. Inf. Syst.,* 22(1), 54-88.

Miquel, M., Beatriz, L., pez, Josep, L., & s De La, R. (2003). A Taxonomy of Recommender Agents on theInternet. *Artif. Intell. Rev.,* 19(4), 285-330.

Mitchell, T. (1997). *Machine Learning.* McGraw-Hill.

Naak, A., Hage, H., & Aïmeur, E. (2008). Papyres: a Research Paper Management System, *IEEE Joint Conference on E-Commerce Technology (CEC 08) and Enterprise Computing, E-Commerce and E-Services (EEE 08)* (pp. 201-208). Crystal City, Washigton, D.C., USA.

Naak, A., Hage, H., & Aïmeur, E. (2009). A Multi-criteria Collaborative Filtering Approach for Research Paper Recommendation in Papyres. Dans G. Babin, P. K., and M. Weiss (Éd.), *The 4th International MCETECH Conference on e-Technologies (MCETECH 2009)* (Vol. LNBIP 26, pp. 25-39). Ottawa, Canada: Springer-Verlag.

Nishikant, K., Jilin, C., John, T. B., Gary, C. F., James, A. S., John, R., *et al.* (2007). Techlens: a researcher's desktop, *Proceedings of the 2007 ACM conference on Recommender systems* (pp. 183-184). Minneapolis, USA.

O'Reilly, T. (2005). *What Is Web 2.0*, http://www.oreillynet.com/

Pazzani, M., J. . (1999). A Framework for Collaborative, Content-Based and Demographic Filtering. *Artif. Intell. Rev.,* 13(5-6), 393-408.

Popescul, A., Ungar, L. H., Pennock, D. M., & Lawrence, S. (2001). Probabilistic Models for Unified Collaborative and Content-Based Recommendation in Sparse-Data Environments, *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*: Morgan Kaufmann Publishers Inc.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). GroupLens: an open architecture for collaborative filtering of netnews, *Proceedings of the 1994 ACM conference on Computer supported cooperative work*. Chapel Hill, North Carolina, United States: ACM.

Resnick, P., & Varian, H. R. (1997). Recommender systems. *Commun. ACM,* 40(3), 56-58.

Reuters Thomson. (2008). *http://www.thomsonreuters.com/content/PDF/scientific/Web_of_Knowledge_factsheet.pdf*:

Salton, G. (1989). *Automatic Text Processing.* Addison-Wesley.

Schafer, J. B., Konstan, J., & Riedl, J. (1999). Recommender systems in e-commerce, *Proceedings of the 1st ACM conference on Electronic commerce*. Denver, Colorado, United States: ACM.

Schein, A. I., Popescul, A., Ungar, L. H., & Pennock, D. M. (2002). Methods and metrics for cold-start recommendations, *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. Tampere, Finland: ACM.

Shardanand, U., & Maes, P. (1995). Social information filtering: algorithms for automating \“word of mouth\&rdquo, *Proceedings of the SIGCHI conference on Human factors in computing systems*. Denver, Colorado, United States: ACM Press/Addison-Wesley Publishing Co.

Smyth, B., & Cotter, P. (2000). A Personalized TV Listings Service for the Digital TV Age. *Knowledge-Based Systems,* 13(2-3), 53-59.

Soboroff, I. M., & Nicholas, C. K. (1999). Combining Content and Collaboration in Text Filtering, *Int'l Joint Conf. Artificial Intelligence Workshop: Machine Learning for Information Filtering*.

Tang, T. Y. (2008). *The design and sutdy of pedagogical paper recommendation.* University of Saskatchewan.

Tang, T. Y., & McCalla, G. (2004). *Beyond learners' interest: personalized paper recommendation based on their pedagogical features for an e-learning system*, Auckland, New Zealand.

Tang, T. Y., & McCalla, G. (2007). A multi-dimensional paper recommender, *Proceedings of the 13th International Conference on Artificial Intelligence in Education (AIED 2007)*. Marina Del Rey, USA.

Torres, R., McNee, S., Abel, M., Konstan, J., & Riedl, J. (2004). Enhancing digital libraries with TechLens+, *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries (JCDL 04)* (pp. 228-236). Tuscon, USA.

Tran, T., & Cohen, R. (2000). *Hybrid Recommender Systems for Electronic Commerce*: AAAI Press.

Vassileva, J. (2004). Harnessing P2P Power in the Classroom. Dans '2004, I. (Éd.), *Intelligent Tutoring Systems, 7th International Conference (ITS 04)* (Springer Berlin / Heidelberg[e] éd., Vol. 3220/2004, pp. 305-314). Alagoas, Brazil: Springer.

Vassileva, J. (2008). Supporting Peer-to-Peer User Communities. *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE* (pp. 230-247).

Wang, Y., & Vassileva, J. (2004). Trust-Based Community Formation in Peer-to-Peer File Sharing Networks, *Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*: IEEE Computer Society.

## URLs

URL, 1 : *IEEExplore*. Récupéré le Feb. de http://ieeexplore.ieee.org/

URL, 2 : *ACM Digital Library*. Récupéré le Feb. de http://portal.acm.org/dl.cfm

URL, 3 : *EndNote*. Récupéré le Feb. de http://www.endnote.com/

URL, 4 : *CiteULike*. Récupéré le Feb. de http://www.citeulike.org/

URL, 5 : *SpringerLink*. Récupéré le Feb. de http://springerlink.metapress.com/

URL, 6 : *BibTeX*. Récupéré le Feb. de http://www.bibtex.org/

URL, 7 : *Google Scholar*. Récupéré le Feb. de http://scholar.google.ca/

URL, 8 : *CiteSeer*http://citeseer.ist.psu.edu/

URL, 9 : *CiteSeerX*. Récupéré le mai 20 de http://citeseerx.ist.psu.edu/

URL, 10 : *Livelink ECM 10*. Récupéré le Feb. de http://www.opentext.com/2/sol-products/sol-pro-llecm10.htm

URL, 11 : *Zotero*. Récupéré le may 2009 de http://www.zotero.org/

URL, 12 : *Documentum*. Récupéré le Feb. de http://www.documentum.com/

URL, 13 : *YAHOO! RESEARCH*. Récupéré le Juin de http://research.yahoo.com/

URL, 14 : *YAHOO! MOVIES*. Récupéré le Juin de http://movies.yahoo.com/