Usagers & Recherche d'Information

Max Chevalier

Mémoire pour l'obtention de

L'HABILITATION à DIRIGER des RECHERCHES

Spécialité informatique

USAGERS

&

RECHERCHE D'INFORMATION

Max Chevalier

Présenté le 24 février 2011 devant la commission d'examen composée de :

Catherine BERRUT	***Professeure, Université Joseph Fourier Grenoble***	**Rapporteure**
Sylvie CALABRETTO	***Professeure, INSA Lyon***	**Rapporteure**
Corine CAUVET	***Professeure, Université Aix-Marseille III***	**Examinatrice**
Claude CHRISMENT	***Professeur, Université Toulouse III - Paul Sabatier***	**Examinateur**
Thérèse LIBOUREL	***Professeure, Université Montpellier II***	**Rapporteure**
Florence SÈDES	***Professeure, Université Toulouse III - Paul Sabatier***	**Examinatrice**
Chantal SOULÉ-DUPUY	***Professeure, Université Toulouse 1 - Capitole***	**Directrice de recherche**

Institut de Recherche en Informatique de Toulouse (UMR 5505)

Université Paul Sabatier (Toulouse III) – 118 route de Narbonne – F-31062 Toulouse cedex 09

« La ***théorie***, c'est quand on sait tout et que rien ne fonctionne.

La ***pratique***, c'est quand tout fonctionne et que personne ne sait pourquoi.

Si la ***pratique*** et la ***théorie*** sont réunies, rien ne fonctionne et on ne sait pas pourquoi. »

Albert Einstein

A Isabelle, Marc et Ivan

Table des Matières

A. INTRODUCTION

A.1 Préambule

L'information peut être considérée aujourd'hui comme un capital (immatériel) important[1] tant pour les organisations que pour les individus eux-mêmes. Cette information permet de se construire ou valider une connaissance concernant différents objets de la vie quotidienne (concurrents, langages de programmation, sports…).

Dans le même temps, on peut souligner que l'information se retrouve partout, dans les intranets, les extranets et particulièrement sur le Web qui n'en finit plus de grandir et qui constitue une source d'information privilégiée. Il n'est donc pas surprenant de constater que le nombre de requêtes sur les moteurs de recherche du Web augmentent de façon significative. Une étude de *comeScore*[2] souligne une augmentation de 41% du nombre de requêtes (durant le mois de juillet 2008 et 2009) sur les principaux moteurs du Web (113,685 milliards de requêtes en 2009).

Pour identifier les informations répondant à ses besoins, l'usager pratique ce que nous définissons comme une activité de **recherche d'information**. Il s'agit d'une activité réalisée par un très grand nombre d'usagers que ce soit dans un contexte de loisirs ou professionnel. Elle se combine couramment avec d'autres activités (diffusion, rédaction d'un document) participant ainsi à un processus de gestion documentaire (ou gestion électronique de documents). Ce processus occupe aujourd'hui une place importante dans l'activité quotidienne des individus.

Cette activité de recherche d'information peut revêtir différentes formes allant de la simple demande à un ami/collègue à l'usage de systèmes informatiques (systèmes de recherche d'information, systèmes hypermédia, systèmes de recommandation…). Dans tous les cas, l'objectif est le même : obtenir l'information utile (la plus adaptée) dans les meilleures conditions (temps, ressources…).

Pour être efficaces, les outils informatiques pour la recherche d'information doivent relever un défi important : faire face à une **hétérogénéité** croissante, surtout sur le Web, des usagers, des informations et des contextes dans lesquels les informations sont recherchées.

Dans ce cadre, une des solutions les plus importantes proposée dans la littérature est sans doute l'application des principes de **personnalisation** dans le domaine de la recherche d'information. Initialement la personnalisation proposée vise à tenir compte de l'hétérogénéité des usagers. Elle repose principalement sur des modèles usagers pour réaliser l'adaptation souhaitée. Plus récemment la personnalisation a intégré une notion plus large de contexte (Figure 1) incluant des dimensions autres que l'usager et que les besoins informationnels (environnement, tâche…).

[1] http://www.observatoire-immateriel.com/IMG/pdf/Capital-Immateriel-N10-11.pdf

[2] http://www.comscore.com/Press_Events/Press_Releases/2009/8/Global_Search_Market_Draws_More_than_100_Billion_Searches_per_Month

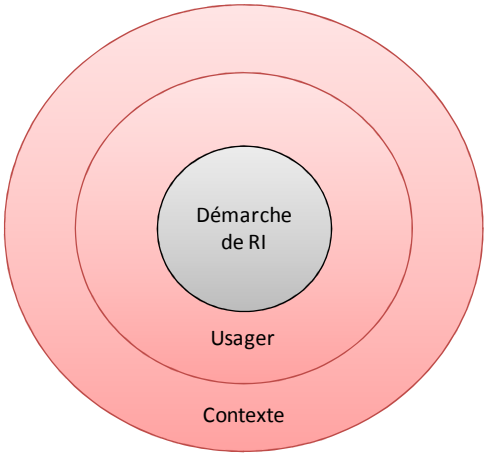


Figure 1 - Évolution de la Recherche d'Information (RI) face à l'hétérogénéité des usagers et de l'environnement (les deux niveaux périphériques sont exploités à des fins de personnalisation)

A.2 Orientation de nos travaux

Dans ce contexte et sur la base des résultats obtenus au cours de ma thèse (Chevalier, 2002), nos travaux de recherche s'intéressent aux différentes activités documentaires (dont fait partie la recherche d'information) et particulièrement aux relations qu'elles peuvent entretenir entre elles.

L'objectif que nous nous sommes fixés est d'améliorer globalement les activités des usagers liées à l'information en tenant compte de l'hétérogénéité précédemment identifiée :

- à un niveau **macroscopique** : nos travaux portent sur l'analyse et l'amélioration globale des activités documentaires parmi lesquelles la recherche d'information joue un rôle essentiel,
- à un niveau **microscopique** : nos travaux se sont focalisés sur une activité documentaire en particulier : l'activité de recherche d'information. L'objectif est d'améliorer cette activité en améliorant plus particulièrement les systèmes de recherche d'information. Les travaux de recherche menés consistent principalement à améliorer les « points d'interaction » entre le système de recherche d'information et l'usager : la formulation des requêtes et la visualisation des résultats de recherche,
- au niveau des **usagers** : nos travaux portent sur la définition de **modèles pour la personnalisation** en Recherche d'Information. Ces modèles correspondent aux deux niveaux de personnalisation présentés dans la Figure 1 (modélisation des usagers, du contexte).

Les travaux que nous avons menés depuis 2002 visent globalement l'amélioration de la relation entre usagers et informations. De part notre expérience dans ce domaine, la **Recherche d'Information** est centrale à nos propositions tout comme la **dimension « usager »**. Enfin, les résultats obtenus et présentés dans ce mémoire sont le fruit de travaux structurés et développés dans le cadre de cinq **thèses** (soutenues ou en cours) et trois stages de **Masters 2** que j'ai co-encadrés.

A.3 Plan du mémoire

Les axes de recherches s'articulent de la manière suivante dans ce mémoire :

- le **chapitre B** présente l'activité de Recherche d'Information. La Recherche d'Information constitue en effet la pierre angulaire de nos travaux. Une présentation des concepts clés est faite avant de souligner l'apport des approches de personnalisation appliquées à la recherche d'information.

- le **chapitre C** présente le premier axe de recherche correspondant au **niveau macroscopique** de la gestion documentaire. Il s'agit d'une **architecture de fédération** des activités documentaires permettant notamment un enrichissement mutuel de ces activités et visant à les améliorer. Cette architecture repose sur un ensemble de **processus** et **d'indicateurs** liés aux activités des usagers. Pour ce faire, nous définissons et exploitons le concept **d'annotation collective**.

- le **chapitre D** présente le deuxième axe de recherche correspondant au **niveau microscopique** de la gestion documentaire c'est-à-dire au niveau des systèmes de recherche d'information. Nous abordons les problématiques de la formulation des besoins et de l'exploitation des résultats de recherche. En réponse à ces problématiques, nous proposons un système de médiation pour la désambigüisation de requêtes, une interface de visualisation des résultats de recherche multi-facettes ainsi qu'une approche permettant de mesurer l'efficacité des interfaces de visualisation des résultats de recherche.

- le **chapitre E** présente l'axe de recherche transversal correspondant à la **modélisation** des ressources (usagers notamment) et du contexte pour la personnalisation en Recherche d'Information.

- enfin, le **chapitre F** présente une vision synthétique des travaux réalisés et souligne notamment la valorisation scientifique qui en a été faite. Le support de la recherche (thèses, M2) et les collaborations que nous avons eues favorisant l'obtention des résultats présentés dans ce mémoire sont également mis en avant.

B. RECHERCHE D'INFORMATION ET PERSONNALISATION

B.1 Recherche d'information : concepts clés

Avant de présenter plus en détails l'activité de Recherche d'Information, nous proposons d'éclaircir au travers de quelques définitions le concept de **Recherche d'Information**.

D'après l'AFNOR la ***Recherche d'Information*** peut se définir comme :

« *Action, méthodes et procédures ayant pour objet d'extraire d'un ensemble de documents les informations voulues. Dans un sens plus large, toute opération (ou ensemble d'opérations) ayant pour objet la recherche, la collecte et l'exploitation d'informations en réponse à une question sur un sujet précis* ».

Une notion proche de la ***Recherche d'Information*** est la ***Recherche Documentaire*** étant définie par l'AFNOR par :

« *Action, méthodes et procédures ayant pour objet de retrouver dans des fonds documentaires les références des documents pertinents. Ensemble des techniques et modalités permettant de sélectionner l'information dans un fonds documentaire structuré en fonction de critères de recherches propres à l'utilisateur*[3] ».

Ces deux notions assez différentes sur l'objectif, se confondent dans les systèmes actuels du fait notamment de la dématérialisation massive des documents et de l'intérêt toujours grandissant du Web. L'utilisateur peut donc indifféremment réaliser ces deux types de recherche et ce de façon transparente. Il y a donc un accès direct au contenant et au contenu. De ce fait, dans ce document, nous ne parlerons que de **Recherche d'Information (RI)** qui englobe ces deux approches.

La plupart des activités quotidiennes font appel à la RI et ce quel que soit le contexte (professionnel, loisirs…) pour combler un manque en information, vérifier voire valider une information… Dans cette section, nous présentons les différentes approches pouvant être mises en œuvre par l'usager pour (re)trouver de l'information : la recherche d'information et l'approche duale de la RI le filtrage d'information. Nous soulignons également l'intérêt pour un usager de posséder des systèmes personnalisés lui permettant d'obtenir des informations plus adaptées.

B.1.1 La recherche d'information

Il y a encore quelques années, pour rechercher des documents dans une bibliothèque, il fallait parcourir les nombreuses notices bibliographiques classées par mots-clés, par titres, par auteurs afin d'identifier la côte de l'ouvrage tant recherché. Aujourd'hui, les systèmes de recherche d'information ou SRI permettent de rechercher des informations selon des

[3] Dans ce mémoire, les termes de « usager » et « utilisateur » sont synonymes. Ils représentent la personne qui recherche de l'information.

critères plus riches tels que l'information même contenue dans les documents (ex : qui était le président de la France en 1975 ?) et non pas uniquement au travers des caractéristiques exogènes des documents.

Nous présentons dans cette section le processus sur lequel reposent les systèmes de recherche d'information. Ce processus est appelé processus en U et peut être schématisé comme le présente la Figure 2.

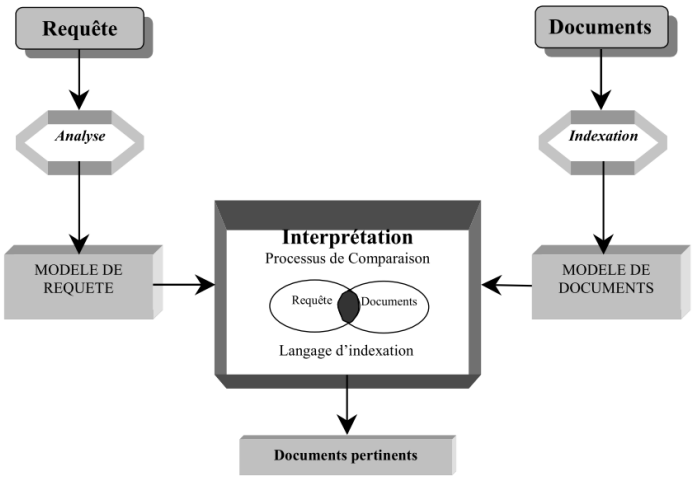


Figure 2 - Le processus en U de la recherche d'information

Dans cette figure, quatre principales étapes se dégagent :

- Le **traitement des documents** qui constituent le fonds documentaire du système. Dès qu'un document est ajouté au fonds documentaire, ce dernier est analysé afin de construire l'équivalent d'une notice bibliographique. Le système identifie donc et extrait les caractéristiques les plus importantes des documents permettant ainsi de répondre aux besoins d'un plus grand nombre d'usagers : phase **d'indexation**. Pour permettre à l'usager de retrouver des documents par rapport à leur contenu, l'indexation d'un document passe par l'indexation de son contenu. Cette indexation, de plus en plus automatisée, peut varier d'un système à l'autre. Les caractéristiques ainsi extraites sont structurées dans un modèle permettant la recherche d'informations répondant à une requête d'un utilisateur,
- La **formulation des besoins** d'un utilisateur, qui consiste à traduire le besoin mental en une *requête* qui sera soumise au système. Cette requête (généralement quelques mots-clés) est analysée et indexée de la même manière que les documents. Un modèle compatible avec celui des documents est alors utilisé pour les requêtes,
- La **recherche** à proprement parler qui consiste à identifier dans le fond documentaire les documents répondant aux besoins de l'usager. Cette phase correspond à un appariement entre le modèle de requête et les modèles de documents. Selon les modèles utilisés, cet appariement donne lieu à une valeur graduelle traduisant l'importance de cet appariement (similarité). Cette similarité la *pertinence système* appelée également RSV (Retrieval Status Value). La *pertinence système* est à mettre en relation avec la *pertinence usager* (Mizzaro, 2007) qui traduit l'intérêt réel d'une information pour un usager,
- La **présentation des résultats** à l'usager. Le système retourne les documents jugés comme les plus pertinents le plus souvent sous la forme d'une liste de résultats

ordonnés par RSV décroissante (la plus élevée correspondant à un appariement optimal).

Du fait que nous ne traitons pas en détails toutes ces thématiques dans ce mémoire, nous renvoyons le lecteur à la bibliographie riche dans le domaine de la RI et particulièrement à (Manning, et al., 2008) et (Baeza-Yates, et al., 1999).

B.1.2 Le Filtrage d'information

Le filtrage d'information ou FI est un processus dual à la recherche d'information comme le souligne (Belkin, et al., 1992). L'objectif de ce processus est d'identifier dans un flux documentaire, les documents correspondant à des centres d'intérêt « stables » d'un usager appelé aussi « profil ». Malone *et al.* dans (Malone, et al., 1987) ont mis en évidence trois modes de filtrage qui peuvent être au besoin combinés :

- le filtrage *collaboratif*. Il se base sur des règles de sélection permettant d'évaluer un document sur la base des jugements de pertinence émis par l'ensemble des utilisateurs du système,
- le filtrage *économique*. Il se base sur les notions de coût et d'intérêt relatifs à la production et la lecture d'un document. Le but est de filtrer l'information pour limiter le coût et maximiser l'intérêt. Le coût et l'intérêt sont représentés par des constantes définies par l'utilisateur,
- le filtrage *cognitif* (ou basé sur le contenu). Il exploite exclusivement le contenu d'un document pour évaluer sa pertinence. Il suppose également que chaque utilisateur opère indépendamment les uns des autres.

Un système de FI identifie sur la base d'une fonction de décision si une information peut être proposée (de façon synchrone ou asynchrone) à l'usager. La connaissance de l'usager ainsi que la fonction de décision qui l'exploite sont des points essentiels dans le filtrage d'information. Or, à l'initialisation du programme, le système ne possède pas une connaissance suffisante pour pouvoir correctement évaluer si une information est pertinente ou non pour un usager. Différents travaux tentent de pallier ce problème appelé « démarrage à froid ». Cet aspect est développé dans la section B.2.2.2.3.1 de ce mémoire.

B.1.3 Dualité entre RI et FI

La dualité entre recherche d'information et filtrage d'information peut être soulignée au travers des critères présentés dans le Tableau 1.

	Recherche d'information (RI)	*Filtrage d'information (FI)*
Besoin en information	Momentané	Permanent
Collection de documents	Statique	Dynamique
Interaction	Très interactif	Peu interactif

Tableau 1 - Dualité entre RI et FI d'après (Belkin, et al., 1992)

Le besoin en information en RI est considéré comme momentané car lié au moment où la recherche est effectuée. A l'inverse, pour le FI, qui repose sur la connaissance de l'usager, les besoins sont considérés plus stables dans le temps.

La collection de documents pour la RI correspond au fonds documentaire composé de documents que le système a préalablement indexé ; il est donc considéré comme statique.

En FI le flux documentaire est plus dynamique car le filtrage peut être appliqué dès l'apparition d'un document dans le système.

D'autre part, la RI, nécessite de la part de l'usager une interaction importante nécessitant la formulation de la requête, le tri des documents et ce à chaque fois que l'usager effectue une recherche. A l'inverse, le filtrage requiert une interaction moindre puisque les informations pertinentes sont filtrées et sont automatiquement soumises à l'usager au cours du temps.

Malgré ces différences, la RI et le FI partagent le même objectif qui est de répondre à un besoin en information d'un usager en lui proposant des documents. La section suivante aborde la transposition de la recherche d'information et du filtrage d'information dans le contexte du Web devenu une source privilégiée d'informations pour toute recherche d'information. Nous soulignons également l'impact des approches collectives (ou sociales) encouragées par les outils Web 2.0 qui permettent à un usager de mettre à disposition des autres usagers des informations favorisant ainsi leurs accès et recherches.

B.1.4 La Recherche d'Information sur le Web

Les techniques de recherche/filtrage d'information ont très vite été adaptées dans le contexte du Web. La recherche d'information a donné naissance aux moteurs de recherche, et le filtrage d'information aux outils de recommandation. En effet, la richesse de contenu de la masse de documents qui composent le Web encourage la plupart des usagers à rechercher l'information sur le Web. Hormis les considérables difficultés de passage à l'échelle et le traitement du contenu des pages Web (peu « propres »), les principes généraux de RI restent les mêmes. Du point de vue de l'usager, une modalité de recherche complémentaire basée sur les hyperliens disponibles dans les pages web est omniprésente : la navigation. La *navigation* consiste à parcourir l'hypermédia composé par les liens hypertextes à la recherche de documents pertinents. La navigation joue un rôle très important dans la RI sur le Web et est complémentaire aux outils de recherche (Ihadjadène, 1999). Ainsi les usagers passent de l'une à l'autre de ces modalités de manière naturelle souvent sans s'en rendre compte.

Cette navigation peut aussi se transformer en « sérendipité », c'est-à-dire qu'elle permet de trouver une information pertinente de façon imprévue, en cherchant autre chose par exemple (Ertzscheid, et al., 2003).

B.1.5 La RI et les évolutions du Web 2.0 : le renouveau social visant l'amélioration de la RI

L'appellation même de « Web 2.0 » ne fait pas consensus. Pourtant tout le monde s'accorde sur l'impact qu'ont eu les évolutions sur l'usage du Web par les usagers. En effet, le passage du web passif au web participatif (ou Web Social) est bel et bien une réalité. Ainsi les usagers ne mettent plus seulement leurs pages Web en ligne, ils donnent leur avis, collectent et partagent de l'information… Un rapport de 2007 (Li, 2007) souligne cette évolution en donnant les grands niveaux d'implication des usagers vis-à-vis de l'information disponible sur le Web. L'auteur de ce rapport présente cette implication sous la forme d'une échelle allant de l'usager « inactif » à celui de « créateur » (Figure 3).

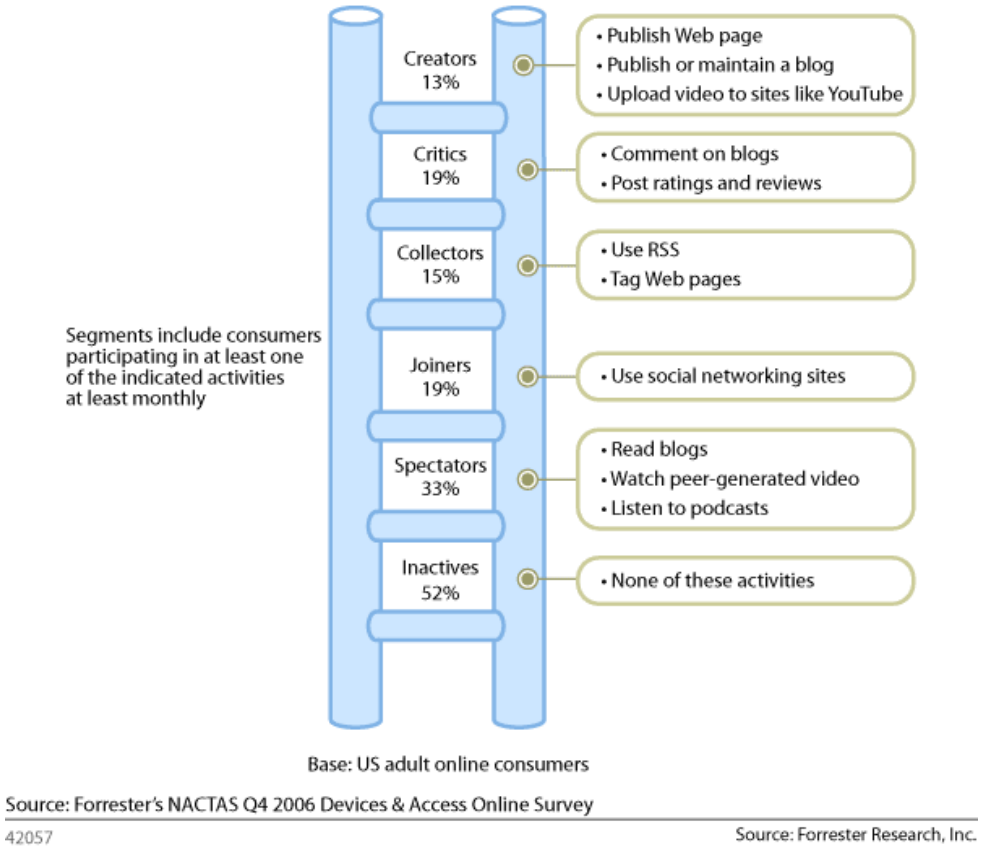


Figure 3 - Implication des usagers (Li, 2007)

Plus son implication augmente, plus l'usager est amené à utiliser des outils adaptés à son activité. La Figure 4 souligne la variété des outils disponibles permettant aux usagers de participer, collaborer et échanger des informations.

Figure 4 - Quelques applications du Web associées à leur objectif principal

Dans le cadre de la RI, il est intéressant de souligner que, grâce à ces applications, les usagers mettent en place de nouvelles formes de RI. Ainsi, outre les classiques moteurs de recherche et outils de recommandation (Montaner, et al., 2003), les usagers s'organisent pour construire de véritables mines d'or d'information favorisant la RI. Pour illustrer ces propos nous présentons, dans la section suivante, au travers de quelques exemples d'applications les différentes formes d'organisation sociale que les usagers mettent en place pour favoriser la RI. Nous étudions ensuite l'impact du Web 2.0 dans les outils de recherche et la façon dont ces outils tentent d'exploiter ces nouvelles mines d'information.

Dans la vie courante, lorsque nous sommes confrontés à un problème, nous faisons souvent appel à autrui. Ce comportement est omniprésent dans le contexte de la RI car bien que les moteurs de recherche soient efficaces, les experts en information restent les humains. Cette vision revient en force dans le contexte de la RI car au bout du compte les experts restent humains. Ainsi, le fait que les usagers puissent collaborer, échanger de façon consciente ou non etc. privilégie la transposition de ce mode de fonctionnement sur le Web. On peut souligner l'émergence d'organisations sociales favorisant la RI par notamment le biais du partage d'information. Nous pouvons qualifier cette organisation de « communautés d'intérêt » (Renner, 2001) car les usagers partagent des centres d'intérêt et contribuent à les faire vivre. Dans le même temps ils construisent des fonds documentaires riches favorisant la RI.

B.1.5.1.1 Plateformes d'étiquetage social

Un premier niveau d'organisation sociale pour la RI peut être réalisé au travers des plateformes d'étiquetage social (« social bookmarking ») tels que *Del.icio.us*[4]. Au travers de ces plateformes, les usagers partagent des références vers des documents qu'ils jugent intéressants. Ces documents sont « indexés » par le biais d'un ensemble de mots-clés choisis librement par les utilisateurs : un ensemble de « tags ». Un tag est en fait une évolution des signets (« bookmark ») qui sont : centralisés, disponibles depuis n'importe quel point d'accès, éventuellement partagés avec les autres usagers et surtout qui permettent un classement multiple des documents si plusieurs tags différents sont associés à un même document. Pour les usagers, ces tags ont une grande importance car ils permettent de retrouver des documents par rapport à ces tags. C'est ainsi que les usagers ont la possibilité de rechercher de l'information par le biais de tags soit au travers d'un système de recherche d'information classique (à base de requête) soit par navigation au travers d'un nuage de tags (Figure 5). Dans un tel nuage, la taille des mots est généralement proportionnelle à l'importance du tag.

Figure 5 - un nuage de tags les plus populaires sur Del.Icio.us (au 5 juillet 2010).

[4] http://delicious.com

Dans ce type de plateforme, le partage d'information peut parfois primer sur l'intérêt personnel. En effet, utiliser des tags inconnus par le commun des mortels ne permettra pas un partage efficace. Ainsi, (Golder, et al., 2006) souligne une tendance à l'uniformisation des mots utilisés comme tags (par exemple l'utilisation du singulier plutôt que du pluriel). Ainsi au sein de ces plateformes peuvent cohabiter des tags « populaires » et d'autres plus individuels décrivant l'intérêt (ou point de vue) d'un individu relatif à un document. Ces plateformes où chacun peut apporter sa contribution peuvent convenir à différents contextes. Ainsi, elles ont été déployées dans les domaines de la recherche scientifique tels que *Connotea*[5] (Lund, et al., 2005) ou encore des entreprises tels que *DogEar* (Millen, et al., 2006). Par le biais de celles-ci, la RI conduit à de meilleures résultats car elle repose sur une indexation manuelle des documents et donc plus proche des usagers.

Dans ces plateformes, une organisation sociale plus ou moins structurée peut exister. En effet, des groupes d'usagers peuvent se former autour des documents indexés. Cette indexation est généralement réalisée par le biais des tags. Le niveau d'accès aux documents peut être privé, limité à un groupe d'usagers ou public. Cette visibilité est fixée par l'utilisateur ayant ajouté des tags à un document. Il existe également dans la plupart de ces plateformes la possibilité de s'abonner à des flux RSS afin de surveiller l'activité d'un usager.

B.1.5.1.2 Plateformes de répertoires partagés

De façon un peu plus structurée, nous pouvons prendre l'exemple des plateformes de partage d'information telles que les « pearl trees »[6] (Figure 6). Dans cette application, on peut voir l'émergence d'une communauté d'intérêt de façon un peu plus précise. L'usager peut créer autant d'arbres qu'il souhaite traduisant la variété de ses centres d'intérêt. Dans chaque arbre, l'usager peut associer des documents (perles) aux nœuds de l'arbre (proches de la notion de répertoire). Le nom d'un répertoire peut être rapproché des tags. La nouveauté consiste en la possibilité de relier et organiser les répertoires ainsi que les documents. Ainsi un usager peut intégrer dans ses propres arbres, des nœuds d'un autre usager. De cette façon, son propre arbre grandit au fur et à mesure à partir des documents collectés par d'autres usagers.

[5] http://www.connotea.org/

[6] http://www.pearltrees.com

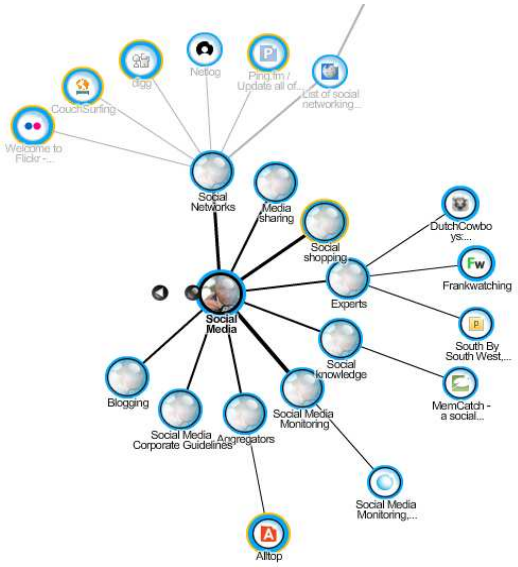


Figure 6 - Un exemple d'arbre de perles concernant les média sociaux (collecté le 10/06/2010))

Dans ces arbres de perles, les relations entre les usagers sont plus explicites bien qu'unilatérale. En effet, utiliser un répertoire d'autrui donne accès au résultat de l'activité de son propriétaire mais ne permet pas d'en modifier le contenu.

B.1.5.1.3 Portails de recherche collaboratifs

De façon encore plus structurée, nous pouvons souligner l'émergence de portails de recherche thématiques. Nous pouvons ainsi citer par exemple les « swickis » proposés l'application Eurekster[7] (Figure 7). L'idée est de permettre à une personne de créer un moteur de recherche personnalisé sur un thème donné. L'originalité de cette démarche est que chacun peut contribuer à ces outils de recherche (en commentant, en réordonnant les résultats etc..) et donc de construire de véritables *outils de recherche communautaires*. L'approche est orientée thématique et les personnes intéressées par un thème peuvent associer ces swickis à leur propre site.

Une initiative assez similaire sur le principe avait été proposée précédemment par Google (*Google SearchWiki*) dans lequel les usagers pouvaient noter et réordonner les résultats. Cette idée a été abandonnée car d'après leur analyse (Dupont, et al., 2010), les usagers apprécient d'indiquer des documents pertinents pour une requête mais n'aiment pas réordonner les résultats. Ainsi, Google propose désormais la possibilité de créer un moteur de recherche personnalisé plus adapté à la demande des usagers au travers de la fonctionnalité *Google Co-op*[8].

[7] http://www.eurekster.com

[8] http://www.google.com/cse/

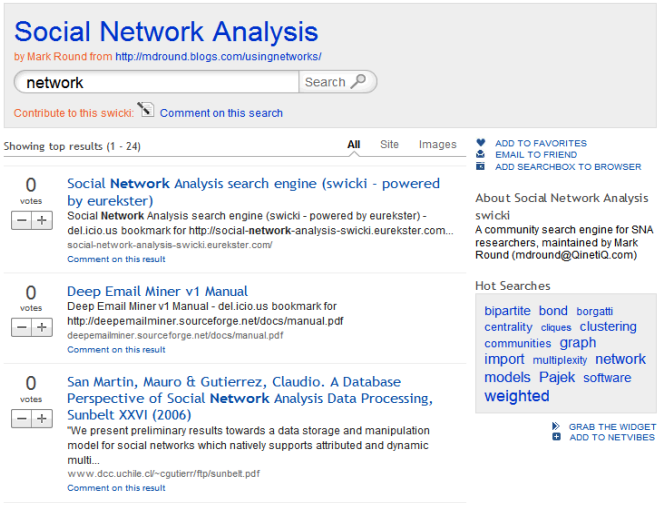


Figure 7 - Exemple de portails de recherche

Ces différentes approches soulignent comment des usagers peuvent se regrouper afin de partager et retrouver des informations plus pertinentes tout en créant de véritables communautés d'intérêt autour de thématiques qu'ils ont en commun. Ainsi, il y a émergence consciente d'une véritable « intelligence collective » qui profite à la RI. Chaque individu peut ainsi profiter de l'activité de groupe pour améliorer ses propres connaissances relatives à un domaine ou pour favoriser la diffusion d'informations autour de ce thème lui conférant ainsi une visibilité accrue.

L'aspect collectif est donc un vecteur d'amélioration de l'activité de RI.

Dans le même objectif d'amélioration que ces approches « sociales », différents travaux se sont plus particulièrement intéressés à faire converger la pertinence système et la pertinence usager. Pour ce faire, un des moyens est de mieux connaître l'usager, ses besoins, ses préférences etc. afin de mieux lui répondre. Les outils deviennent ainsi de plus en plus personnalisés. Cette nouvelle dimension est présentée dans la section suivante.

B.2 Recherche d'Information et Personnalisation

La **personnalisation** dans un contexte général est définie[9] comme :

- **1)** « Action de donner (à quelque chose qui existe en de très nombreux exemplaires, à un objet de série) un caractère singulier, personnel (en particulier en le décorant) »,
- **2)** « Adapter à chaque cas particulier, à chaque personne ». Dans cette définition, ***adapter*** signifie « appliquer, ajuster ou appliquer convenablement, mettre en accord, approprier ».

D'après ces définitions on comprend bien l'objectif de la personnalisation en RI : obtenir des résultats propres à chaque usager en fonction de ses besoins. Selon ce point de vue, la standardisation de la RI ne peut pas être la solution permettant à tous de trouver l'information qu'ils recherchent dans les meilleures conditions. Un ensemble de

[9] Source : *Le Petit Larousse Illustré*, 1996.

caractéristiques humaines doivent être mises en jeu dans le processus de RI pour améliorer la qualité du résultat de recherche obtenu. Cette idée de personnalisation de la RI n'est pas récente. Déjà en 1992, les approches de filtrage d'information (Belkin, et al., 1992) abordaient la problématique de la prise en compte de la dimension usager. Puis en 1997, Brookes dans (Brookes, 1997) soulignait l'importance des approches de recherche d'information orientées *cognition (centrée sur l'usager)* venant en complément des approches classiques de recherche d'information dites orientées *système*.

Afin de souligner l'intérêt des approches de personnalisation dans la RI, nous présentons tout d'abord dans cette section la RI du point de vue de l'usager (aspects cognitifs). Nous soulignons ensuite les principales démarches de personnalisation mises en œuvre pour la RI.

B.2.1 La RI du point de vue de l'usager (aspects cognitifs)

Lors de l'activité de RI, l'usager doit s'impliquer fortement s'il souhaite obtenir des résultats optimaux. D'un point de vue cognitif, il met essentiellement en œuvre, dans les différentes phases de la RI (Chevalier, et al., 2009) :

- ses capacités (physiques, mentales, psychiques…),
- ses connaissances (ou expertises) (Marchionini, 1997), (Hölscher, et al., 2000), (Ciaccia, 2008), :
    - ses **connaissances pratiques** (langages, outils informatiques, outils d'interrogation, …) qui désignent les différentes compétences et connaissances nécessaires au bon usage des outils disponibles,
    - ses **connaissances du domaine** de recherche qui désignent les connaissances dont dispose l'utilisateur sur le thème de recherche,
    - ses **connaissances du processus de RI** définies dans (Marchionini, 1997) comme « connaissances générales qu'une personne a sur les facteurs relatifs à la recherche d'information et leurs habiletés et attitudes pour exécuter le processus informationnel ». Il s'agit des connaissances permettant d'établir la façon d'aborder et de résoudre le problème de RI.

Nous illustrons dans les sections suivantes le rôle et l'impact de ces éléments dans le processus de RI.

B.2.1.1 L'usager dans le cadre de la navigation

Malgré sa facilité apparente, la navigation n'est pas évidente. La principale difficulté de la navigation se situe au niveau de l'effort cognitif requis comme cela est souligné dans (Agosti, et al., 1996), (Baeza-Yates, et al., 1999). En effet, un effort important est consenti par l'usager pour construire une représentation mentale de l'hypertexte dans lequel il navigue (il mémorise tous les documents qu'il visite ainsi que les liens qu'il parcourt). Il peut alors se produire une surcharge cognitive lorsque l'usager n'arrive plus à mémoriser la structure de l'hypertexte dans lequel il navigue (concept du « *lost in the cyberspace* » c'est-à-dire qu'il ne sait plus où il en est). Par ailleurs, pour chaque document visité l'usager doit évaluer s'il répond ou non à ses besoins. Ainsi, la réussite de sa navigation est fortement corrélée à son niveau de connaissance du domaine *de recherche* (est-ce que l'usager connaît ce qu'il cherche par exemple en termes de contenu ou de champ lexical ?). Une autre connaissance est également impliquée dans la navigation. Il s'agit de la connaissance pratique. L'usager sait-il comment est construit une URL comment l'interpréter, comment manipuler les outils mis à disposition. Si tel est le cas il pourra effectuer des recherches plus fructueuses en moins de temps. Enfin, même si cela n'apparaît de prime abord, la navigation

demande une activité cognitive importante pour définir notamment le plan d'actions à réaliser. La *connaissance du processus de RI* est dans ce cadre mise en jeu.

B.2.1.2 L'usager dans le cadre de la recherche et filtrage d'information

Du point de vue de l'usager, il est important de différencier les deux types d'approches puisqu'elles ne demandent pas le même effort notamment en termes d'interaction (Tableau 1, p. 15).

B.2.1.2.1 Recherche d'information (ou « requêtage ») ou systèmes de recherche d'information

La recherche d'information implique l'usager de manière importante notamment au niveau de l'interaction. En premier lieu, lors de la **formulation des besoins**, celui-ci doit transformer ses besoins en une *requête* exploitable par la machine. Les connaissances du domaine de recherche et du processus de recherche sont importantes pour la construction de cette requête. En effet, le choix et le nombre de termes utilisés conditionnent fortement les résultats de recherche. Des termes trop généraux ne permettent pas de cibler suffisamment les besoins et se traduisent par la restitution d'un nombre de documents trop important pour que ceux-ci soient humainement exploitables. A l'opposé, l'utilisation de termes trop spécifiques induit la restitution d'un nombre de documents trop limité, voire nul. Ceci est principalement dû au fait que les termes choisis par l'utilisateur ne correspondent pas forcément à ceux utilisés dans les documents recherchés alors que les techniques traditionnelles d'indexation n'extraient que le vocabulaire des auteurs (problème de synonymie). Quant au nombre de termes utilisés, celui-ci est relativement faible car, en moyenne, moins de trois mots sont utilisés pour formuler une requête (Silverstein, et al., 1998), (Jansen, et al., 2000), (Spink, et al., 2002). Ce nombre de termes dans les requêtes, sur le Web notamment, ne devrait pas évoluer radicalement dans les prochaines années.

L'implication de l'usager ne s'arrête pas là : après restitution des résultats, l'usager doit les analyser pour identifier ceux qui répondre à ses besoins, **il doit exploiter les résultats**. Sa connaissance du domaine associée à son savoir-faire (à sa connaissance pratique) sont encore mises à contribution afin d'identifier les documents répondant à ses besoins par exemple. Cependant, plus le nombre de documents à juger est important, plus la charge cognitive est importante (charge de traitement essentiellement mentale). L'usager doit alors mettre en œuvre une démarche mentale pour identifier et sélectionner les documents répondant à ses besoins. Pour ce faire il doit par exemple lire la plupart du temps un à un les documents retournés afin d'évaluer leur pertinence réelle. Cette charge cognitive est très importante notamment sur le Web et se traduit généralement par une surcharge induisant soit un abandon de la recherche, soit plus généralement une exploration limitée à la première page de documents retournés (environ 10 ou 20 documents).

B.2.1.2.2 Filtrage d'information

Pour l'usager, les outils basés sur le filtrage d'information sont plus « confortables ». En effet, après avoir initialisé le système (soit en formulant ses besoins, soit en donnant des exemples de documents l'intéressant), l'usager recevra automatiquement de nouveaux documents proposés par le système. L'interaction est donc plus limitée. Qui plus est, le nombre de documents retournés est normalement plus restreint car les outils « filtrent » les résultats en fonction des besoins des usagers. L'usager doit cependant toujours impliquer ses connaissances afin d'évaluer la pertinence des documents proposés. En conclusion, en filtrage d'information les problématiques liées à l'usager sont sensiblement les mêmes qu'en recherche d'information mais semblent être impliquées de façon plus limitée.

B.2.2 Personnalisation en RI

Afin d'améliorer l'activité de RI, les systèmes doivent donner des résultats répondant au mieux aux besoins de l'usager. Pour cela, il est nécessaire que la pertinence système (vision que le système à de la pertinence des documents recherchés) tienne compte notamment des caractéristiques individuelles de ces usagers (connaissance du domaine, préférences…). En effet, pour que la pertinence système puisse se rapprocher de la pertinence usager, il conviendrait d'apprendre à mieux « connaître » l'usager et les besoins qu'il a vis-à-vis de l'information. Pour ce faire, il est important que le système intègre un spectre de caractéristiques sur l'usager le plus large possible lui permettant de personnaliser les résultats fournis.

Pour arriver à ce but, deux stratégies de personnalisation sont envisageables :

- le **paramétrage** ou « **customization** » qui permet à l'usager de configurer les systèmes relativement à des options (ou préférences) intégrées initialement par les concepteurs du système (par exemple pour Google : la langue de l'interface, la langue de recherche, le nombre de résultats par page…). Cette stratégie requiert une démarche active de l'usager puisqu'il est à l'origine de cette configuration. Cette dernière est cependant limitée aux seules options proposées par le système,
- le **profilage** (Cho, et al., 2002) qui permet au système de construire une connaissance de l'usager qu'il acquiert principalement au travers de l'interaction avec l'usager.

Dans la suite de ce mémoire, nous nous concentrerons uniquement sur **l'adaptation** des systèmes plutôt que sur le paramétrage car il offre le plus de possibilités en termes de personnalisation. Du point de vue informatique, le processus d'adaptation peut être résumé en trois phases principales :

- La **collecte d'information** : le système collecte un ensemble d'informations relatives à l'activité de l'usager. Ces informations peuvent être soit fournies au système de façon explicite (l'utilisateur associe une note à un document) soit de façon implicite (le système mesure par exemple le temps de lecture d'un document). Une attention particulière doit être portée à ces informations car sur la base de celles-ci le système peut se construire une représentation de l'usager. Dans ce cadre, les systèmes à base de traces (Laflaquière, et al., 2006) peuvent être exploités. Une des particularités de ces systèmes réside dans le fait qu'ils mettent en évidence les différents niveaux d'interprétation de ces traces,

- **L'analyse et la synthèse** des informations collectées : le système nettoie, interprète les informations qu'il possède pour se constituer une connaissance de l'usager. Cette connaissance peut être consommée immédiatement soit sauvegardée et maintenue au cours du temps. Cette connaissance de l'usager sauvegardée par le système au sein d'un modèle correspond à ce que l'on nomme ***profil***. Les problématiques liées aux profils sont exposées plus précisément dans la section B.2.2.2,

- L'**adaptation** à proprement parler : sur la base de la connaissance qu'il détient sur un usager (ou un groupe d'usagers), le système va pouvoir prendre une décision ou appliquer un traitement spécifique tel que la réorganisation de résultats de recherche.

Il est important de souligner le fait que ces différentes étapes ne sont pas nécessairement synchrones.

B.2.2.1 La collecte d'informations pour la RI adaptative

Pour caractériser au mieux l'usager, les systèmes adaptatifs pour la RI collectent des informations au travers de sources d'évidences et notamment des sources provenant du Web 2.0. Ces sources reposant sur des informations générées par les usagers (« user-generated content ») représentent un volume de données très important.

Les sources tangibles suivantes peuvent ainsi être soulignées :

- **l'usager** lui-même, qui peut fournir consciemment des informations au système (par exemple : son nom, son prénom, ses centres d'intérêt…) ;
- **l'historique des requêtes**, qui mémorise la chronologie des besoins en information de l'usager. Certains travaux ont également cherché à conserver non seulement les requêtes des usagers passées mais aussi les résultats renvoyés par les outils de recherche (Raghavan, et al., 1995), (Klink, 2004),
- **l'historique des navigations**, qui mémorise la trace des documents visités par l'usager durant sa recherche (Wexelblat, et al., 1997), (Jaczynski, et al., 1998), (Egyed-Zsigmond, et al., 2003) ; Certaines approches exploitent un historique de navigation enrichi. Ainsi (Danilowicz, et al., 2002), (Arezki, et al., 2004) complètent l'historique par les termes extraits du contenu des documents consultés par l'usager,
- **les jugements de pertinence** dont une revue de littérature est proposée dans (Ruthven, et al., 2003). Ils représentent un jugement de valeur sur un document pour un usager. Ces jugements peuvent être explicites (l'utilisateur affecte une note à un document) ou extraits implicitement de l'observation des actions sur les résultats de recherche notamment (Joachims, 2002), (White, et al., 2002), (Jung, et al., 2007), (Martins, et al., 2009). Certaines études soulignent l'intérêt de combiner ces deux types de collectes pour les jugements de pertinence afin d'améliorer la qualité des résultats de recherche (Quiroga, et al., 1999),
- les annotations associées par l'usager aux documents. Différentes formes d'annotations peuvent être soulignées :
  - les **annotations textuelles** que peut laisser l'usager sur les documents qu'il visite. Ces annotations sont d'une grande valeur ajoutée et peuvent être considérées du point de vue RI comme un complément du contenu des documents eux-mêmes (Agosti, et al., 2005),
  - les **signets** qui mémorisent les URL des documents que l'usager souhaite conserver,
  - les **tags**, que nous avons décrits précédemment. Ils peuvent être de bons indicateurs quant aux centres d'intérêt de l'usager (Carman, et al., 2008),
- **l'organisation des ressources** conservées par l'usager (signets, fichiers, mails, ...) dans son espace personnel d'information noté EPI (Jones, et al., 2005), (Jones, 2007),
- le **réseau social** de l'usager, modélisant le graphe de ses accointances (Kautz, et al., 1997), (Goecks, et al., 2004).

Pour permettre au système de réaliser l'adaptation nécessaire de façon optimale, les informations exploitées pour construire la connaissance relative à l'usager peuvent être soit

positives (étant pertinentes pour l'usager) soit **négatives** (informations non pertinentes) (Hoashi, et al., 2000), (Widyantoro, 1999). Dans la plupart des cas, seules les informations « positives » sont exploitées. Cependant, afin d'avoir une connaissance plus précise de l'usager et de ses besoins, il peut être également utile de prendre en compte les informations « négatives ». L'utilisation des informations négatives conjointement aux informations positives permet en particulier de lever certaines ambiguïtés (Boughanem, et al., 2001). Cette qualité positive ou négative des informations peut être déduite notamment des jugements de pertinence (Kelly, et al., 2003).

Toutes les informations que collecte le système à l'état brut sur un usager (ou un groupe d'usagers) doivent ensuite être épurées, traitées, et synthétisées de façon à pouvoir être exploitées à des fins d'adaptation. Elles pourront ensuite être directement exploitées par le système, ou être conservées pour construire une connaissance à plus long terme de l'usager : *le profil usager*.

B.2.2.2 Notion de profil usager

Le profil de l'usager (ou encore modèle usager) est une représentation de l'usager au sein du système. Grâce à ce profil, le système possède une connaissance sur l'usager grâce à laquelle il peut lui proposer des réponses plus adaptées.

Tout comme le soulignent les auteurs dans (Amadieu, et al., 2009), pour un modèle usager, il faut décrire le « quoi » c'est-à-dire ce qui doit être représenté et le « comment » c'est-à-dire comment représenter le « quoi » au sein du modèle usager. Nous suivons cette logique en présentant dans un premier temps le « quoi », c'est-à-dire les dimensions classiques prises en compte dans un profil dans le contexte de la RI, pour ensuite présenter le « comment » soulignant la façon usuelle de modéliser concrètement un profil au sein du système (représentation interne).

B.2.2.2.1 Les dimensions d'un profil usager pour la RI

Dans le cadre de la RI, un profil était à l'origine centré plus particulièrement sur les besoins de l'usager (Korfhage, 1997) qui est l'une des pierres angulaires de la RI. Plus généralement, aujourd'hui, le profil en RI est devenu multidimensionnel permettant au système d'avoir une vision plus complète de l'usager lui permettant d'atteindre l'objectif de personnalisation attendu. Dans (Lainé-Cruzel, 1999), un modèle usager peut être vu comme un regroupement d'informations décrivant « qui » est l'usager (par exemple, niveau de connaissances dans le domaine), « ce qu'il veut » (par exemple, volume d'informations attendu) et « ce qu'il fera » de l'information (par exemple, tâche qui implique sa recherche).

L'analyse des travaux relatifs à la personnalisation en RI révèle l'inexistence d'un modèle usager de référence pour la RI. Chaque système définit et utilise donc son propre modèle de l'usager qui est en adéquation directe avec ses besoins spécifiques d'adaptation. Cependant, malgré l'absence de standard, le modèle proposé dans le projet APMD peut être considéré comme l'un des plus complets.

De façon plus précise, Le projet APMD (Bouzeghoub, et al., 2005) a eu pour objectif d'identifier les dimensions de l'usager à prendre en compte dans un système de RI afin de proposer un accès personnalisé aux informations. Ce travail est le fruit d'une synthèse des différents travaux de la littérature. Cinq principales dimensions ont été définies :

- les informations relatives aux domaines d'intérêt,
- les informations personnelles,
- les informations relatives à la qualité attendue (des sources, des informations…),
- les informations relatives à la livraison (comment l'usager souhaite obtenir les résultats),
- les informations relatives à la sécurité.

Quelles que soient les dimensions du profil usager considérées dans le modèle, ce dernier doit ensuite être implanté au cœur du système.

B.2.2.2.2 Dimensions d'un profil : aspect stockage (représentation interne)

Afin de rendre ces modèles exploitables du point de vue du système, différentes stratégies peuvent être utilisées. Chaque dimension du modèle peut être implantée sous la forme d'un ensemble de variables pouvant être typée ou traduite dans un modèle de représentation spécifique selon les besoins d'adaptation.

La forme la plus simple que peuvent prendre les variables est la forme « attribut = valeur ». Cette forme simple ne permet cependant pas de mémoriser des informations plus structurées. Elle est très utilisée dans les profils peu structurés et principalement dédiées à des valeurs uniques (ex : le nom, le prénom, l'âge, un identifiant…).

Un des modèles sans doute le plus couramment utilisé en RI est le modèle vectoriel (Salton, 1971) puisqu'il peut être utilisé pour représenter un besoin ou un centre d'intérêt (Goker, et al., 1991), (Armstrong, et al., 1995), (Korfhage, 1997), (Danilowicz, et al., 2002), (Somlo, et al., 2003), (Jeon, et al., 2008). L'intérêt d'un tel modèle pour un profil est qu'il est assez simple à mettre en œuvre (liste d'objets éventuellement pondérés) et est associé à des mesures connues comme par exemple pour calculer des distances entre les vecteurs par le biais des formules usuelles comme la formule du cosinus. Ce modèle est également très utilisé dans le cadre de l'indexation des documents (Baeza-Yates, et al., 1999) permettant ainsi de calculer directement l'appariement entre une variable du profil et des documents.

D'autres modèles permettent d'extraire des relations entre les informations collectées (par exemple les historiques) : matrices (Balfe, et al., 2005), séries temporelles (Trousse, et al., 1999), versions d'objets complexes (Hubert, et al., 2007), graphes (Daoud, et al., 2009), classifieurs hiérarchiques (Chevalier, 2002), (Huete, et al., 2008), systèmes de raisonnement à partir de cas (Jéribi, et al., 2002).

D'autres modèles plus orientés « sémantique » ont également été proposés : réseaux sémantiques (Stefani, et al., 1998), (Arezki, et al., 2004), hiérarchies de connaissances (Berisha-Bohe, et al., 2007), hiérarchies de concepts ou ontologies (Pretschner, et al., 1999), (Nanas, et al., 2003), (Sieg, et al., 2007), (Stan, et al., 2008).

Enfin, il est à noter que bien d'autres modèles basés notamment sur l'apprentissage tels que les modèles issus des domaines du Data Mining (Tufféry, 2010) et plus particulièrement

du Web usage Mining (Liu, 1998) ou (Markov, et al., 2007), d'apprentissage automatique (Cornuéjols, et al., 2002) sont également appropriés et peuvent être utilisés en RI.

Cette liste, qui ne vise pas l'exhaustivité, souligne la diversité des possibilités offertes aux concepteurs des systèmes adaptatifs en RI. Un panel de modèles s'offre à eux : depuis des modèles simples à implanter et à manipuler jusqu'aux modèles structurés incluant une grande plus-value sémantique mais au prix d'un traitement plus lourd. Cependant, malgré toute la richesse et le soin apporté au profil, les systèmes adaptatifs souffrent de trois problèmes majeurs qui sont l'instanciation (ou « démarrage à froid »), la prise en compte du temps dans l'évolution du profil ainsi que le respect de la vie privée.

B.2.2.2.3 Problématiques liées aux profils usagers

B.2.2.2.3.1 Démarrage à froid des systèmes

Les systèmes construits sur la base d'une connaissance de l'usager ne peuvent réaliser une adaptation optimale qu'à partir du moment où cette connaissance est suffisante. Le problème est que lors de l'initialisation, dans le cas d'un nouvel usager par exemple, le contenu du profil initial est généralement vide ne permettant pas d'évaluer par exemple la pertinence des documents pour cet usager. Ce problème est nommé problème de « démarrage à froid ».

Il existe différentes stratégies qui peuvent être combinées pour limiter ce problème :

- La **saisie manuelle** : l'usager donne les informations qui permettront au système de construire un profil initial minimal,

- Les **exemples** : l'usager fournit au système un exemple d'informations qui correspondent à celles qu'il recherche voire qu'il ne recherche pas. Ces informations sont alors traitées comme s'il s'agissait d'informations « classiques » (jugement de pertinence) et sont alors synthétisées afin de construire un profil initial,

- Les **stéréotypes** (Kay, 1994), (Shapira, et al., 1997) : un stéréotype est un profil prédéfini et valorisé correspondant à un groupe/catégorie d'usagers. L'intérêt d'utiliser un stéréotype est qu'il permet au système d'avoir un profil initial valorisé dès l'instant où l'usager est rattaché à au moins l'un d'entre eux. Initialement, l'usage des stéréotypes était exclusif, c'est-à-dire qu'il n'existait pas de profil individuel pour un usager rattaché à un stéréotype. Des travaux proposent une évolution de cet usage en combinant des stéréotypes et des profils usagers « classiques ». Ainsi, le profil usager pourra être initialisé aux valeurs du stéréotype auquel l'usager est rattaché pour ensuite évoluer naturellement (Ardissono, et al., 2001), (Nguyen, et al., 2006).

B.2.2.2.3.2 Notion de temps dans le profil

Un profil usager peut être ou non sensible au temps. S'il est **statique**, il ne peut pas évoluer dans le temps. Cela n'exclut pas que l'usager par exemple peut ponctuellement modifier tout ou partie de son profil. A l'opposé s'il est **dynamique**, ce qui est majoritairement le cas, il évolue dans le temps pour que la connaissance que possède le système sur l'usager suive l'évolution de l'usager lui-même. Il est en effet important que cette connaissance soit la plus fidèle possible pour obtenir des résultats optimaux et ainsi

éviter de réaliser des adaptations erronées (Chevalier, et al., 2008b), (Chevalier, et al., 2009b).

Dans le cas d'un profil dynamique, il est important de s'interroger sur la validité des informations utilisées pour construire le profil. En effet, un profil est généralement incrémental c'est-à-dire qu'une évolution s'ajoute (par le biais d'une fonction d'ajout ou de fusion) au modèle existant. Pourtant, le but étant de suivre l'évolution de l'usager, il est important de considérer l'âge des informations pour moduler leur importance au sein du profil. Les principales approches utilisées dans ce cadre sont :

- L'usage d'une **fenêtre temporelle** (Widmer, et al., 1996). Le système ne considère que les informations situées dans cette fenêtre pour générer le profil. Toutes les autres ne sont pas prises en compte,

- L'usage d'une **fonction d'oubli graduelle** (Koychev, 2000). Le système module l'importance en fonction de leur âge, les informations les plus récentes étant considérées comme les plus importantes.

Les approches précédentes impactent fortement le profil usager ainsi obtenu privilégiant les informations récentes par rapport aux informations plus âgées. Pourtant il peut être tout à fait pertinent de faire cohabiter ces deux types d'informations en privilégiant les informations liées à l'activité courante de l'usager et les informations « à plus long terme » concernant l'usager. Une solution est l'usage de deux types de profils pour le même usager : c'est-à-dire un profil **à court terme** et un profil à **long terme**. Cette distinction permet au système de réagir à des besoins émergents (à court terme) tout en maintenant des besoins récurrents (à long terme). Ainsi, le système peut réagir différemment s'il considère une session de recherche ou s'il considère les besoins de l'usager récurrents dans le temps. Les informations à long terme sont le plus souvent déduites des informations à court terme (Mizzaro, et al., 2002), (Sugiyama, et al., 2004).

B.2.2.2.3.3 Respect de la vie privée

Le respect de la vie privée autour des profils est un élément important mais complexe sur lequel peut reposer l'acceptabilité d'un système. Ces problématiques peuvent être traitées sur la base du volet légal, éthique ainsi qu'individuel (demandes de la part des usagers) (Schreck, 2001).

Différentes solutions et préconisations sont développées notamment dans (Schreck, 2001) et (Kobsa, 2007) pour garantir au maximum le respect de la vie privée dans les systèmes adaptatifs.

B.2.2.3 Applications de la personnalisation pour la RI

Dès l'instant où le profil usager est exploitable, le système peut mettre en place des processus d'adaptation visant à améliorer l'accès à l'information pour l'utilisateur final. Cette section brosse les principales adaptations réalisées dans le cadre de la RI.

B.2.2.3.1 Reformulation de requêtes

La formulation de requête est une étape cruciale dans le processus de recherche d'information. Afin d'aider l'usager à mieux formuler ses besoins, différentes stratégies d'adaptation peuvent être mises en place : l'expansion de requêtes et la réinjection de pertinence. La personnalisation de la requête repose essentiellement sur la **réinjection de pertinence**. L'approche est itérative et repose sur les jugements de pertinence (ou de non

pertinence) dont dispose le système. Par le biais de ces jugements de pertinence le système peut pondérer les termes de la requête initiale, soit ajouter (ou supprimer) des termes pour former une nouvelle de requête (Rocchio, 1971), (Salton, et al., 1990), (Soulé-Dupuy, 2001), (Vinay, et al., 2005).

La reformulation de requête peut également reposer sur les documents que possède l'usager dans son espace personnel d'information (EPI) (Bottraud, et al., 2004).

B.2.2.3.2 Personnalisation des résultats de recherche

Une autre application classique de l'adaptation en RI consiste à personnaliser le résultat de recherche et les documents renvoyés à l'usager en fonction de ses besoins.

Une des approches les plus classiques est de **réordonner et adapter la liste de résultats de recherche** (Pretschner, et al., 1999), (Christine, et al., 1999), (Balfe, et al., 2005), (Speretta, et al., 2005), (Zhuang, et al., 2006), (Chakrabarti, 2007), (Gong, et al., 2007), (Tamine, et al., 2007), (Mylonas, et al., 2008), (Biancalana, 2009) :

- soit en appliquant un post-traitement sur la liste de résultats renvoyés par l'outil de recherche pour modifier la position des documents dans la liste de résultats ou filtrer les résultats pour ne conserver que ceux qui correspondent aux besoins de l'usager,
- soit en adaptant directement la fonction d'appariement utilisée pour adapter le score des documents retrouvés. Pour cela, la fonction d'appariement peut intégrer par exemple les informations relatives au réseau social de l'usager (Kirchhoff, et al., 2008). Dans la même orientation le moteur de recherche *Nomao*[10] exploite les jugements de pertinence des membres du réseau social virtuel de l'utilisateur (*facebook*) pour moduler en partie la fonction d'appariement.

Enfin, les résultats de recherche peuvent être **enrichis,** fournissant à l'utilisateur des informations complémentaires sur le document en soulignant, par le biais d'annotations notamment, les parties du document pertinentes (Marshall, 1998) ou en ajoutant des pointeurs vers d'autres documents (Schilit, et al., 1998). Ce type d'approches est fréquemment utilisé dans le cadre de la navigation sur le Web.

B.2.2.3.3 Navigation

Les applications de la personnalisation dans le cadre de la navigation sont essentiellement proposées autour des outils de recommandation. L'objectif principal des travaux consiste à indiquer à l'utilisateur des liens vers des documents répondant au mieux à leurs besoins. Nous pouvons citer dans ce cadre *Letizia* (Lieberman, 1995) et *WebWatcher* (Armstrong, et al., 1995) qui indiquent à l'usager les liens les plus pertinents dans les documents que l'utilisateur visite. *Broadway* (Jaczynski, et al., 1998) repose quant à lui sur les navigations antérieures des différents usagers du système pour proposer à l'usager (par anticipation) des documents répondant à ses besoins. Pour aider l'usager à distinguer les documents pertinents, *Syskill*&*Webert* (Pazzani, et al., 1996) propose d'indiquer à l'usager le fait que le document qu'il visite est pertinent ou non par rapport à ses centres d'intérêt.

[10] http://www.nomao.com

B.2.3 Aspect collectif dans la personnalisation de la RI

Nous avons souligné dans la section B.1.5 le fait que l'aspect collectif était important pour améliorer la relation entre les usagers et l'information disponible. Cet aspect collectif permet de faciliter la démarche de RI de chaque usager. Du point de vue des systèmes informatiques, nous avons souligné différentes applications de la personnalisation qui avaient pour objectif d'améliorer la démarche de RI d'un usager donné. Cette personnalisation peut prendre en compte l'aspect collectif pour davantage améliorer les résultats de recherche. L'importance de ces approches ont été développées dans (Cabanac, et al., 2008c).

Cet aspect collectif se développera dans les approches de personnalisation de la démarche de RI au même rythme que la croissance des sources de données type Web 2.0. Un exemple révélateur est l'usage des plateformes de réseaux sociaux virtuels (Kirchhoff, et al., 2008) ou encore l'usage des tags (Biancalana, 2009) pour la personnalisation de requêtes.

B.3 Bilan

Ce chapitre présente le cœur de notre étude : le contexte de la RI. Nous avons présenté les différentes démarches possibles permettant à un usager de trouver et retrouver de l'information répondant à ses besoins. Nous avons également souligné l'effort que ces utilisateurs mettent en place pour collaborer afin que cette recherche d'information soit plus aisée.

Malgré tout, ce processus, du point de vue de l'usager, peut se révéler plus compliqué que prévu car la réussite d'une recherche repose pour une grande part sur les compétences et connaissances de l'usager. Nous avons identifié les principales difficultés rencontrées par les usagers lors de l'utilisation de systèmes destinés à la RI que ce soit au cours de la navigation, de l'utilisation des systèmes de recherche d'information ou des systèmes de recommandations. Les approches de personnalisation sont donc privilégiées dans ce cadre pour améliorer la pertinence système afin que celle-ci se rapproche le plus possible de la pertinence utilisateur. Nous avons également étudié l'aspect collectif (du point de vue des outils destinés à la RI) et mis en évidence le fait qu'il permet d'améliorer la RI.

Cependant, la RI n'est généralement pas une fin en soi. Elle s'intègre dans un processus plus large de réalisation d'une tâche dans lequel un ensemble d'activités s'organisent. Nous allons donc repositionner la RI dans un contexte plus général : celui de la **gestion documentaire** (gestion électronique de documents) dont l'objectif est de favoriser la création, la diffusion de documents... Nous présentons la GED au travers de ses différentes activités documentaires dont la RI fait partie.

Dans ce cadre, la structuration de nos travaux vise **l'amélioration de l'activité globale de gestion documentaire** des usagers. Nous traitons ce problème au travers du spectre de la RI qui joue, de notre point de vue, un rôle central dans ces activités.

Dans cet objectif, les éléments de réponse que nous proposons sont abordés dans les chapitres suivants :

- Le chapitre **C** s'intéresse au **niveau macroscopique** de la gestion documentaire. Ce chapitre traite de la relation qui existe entre les différentes activités documentaires. En effet, nous le soulignons dans le chapitre suivant, les activités documentaires sont indépendantes et n'ont pas réellement d'interaction entre elles. L'objectif de nos propositions est de permettre **l'enrichissement mutuel des différentes activités documentaires** pour améliorer de façon globale la gestion documentaire. L'enrichissement proposé repose sur l'activité de

recherche et l'activité d'annotation collective qui peuvent être considérées comme des vecteurs d'intégration des différentes activités. Dans le même temps, nous proposons **deux indicateurs basés sur les activités** des usagers. Ces indicateurs sont exploités notamment dans les différents processus favorisant l'interaction entre les différentes activités et qui constituent la base de l'architecture proposée,

- Le chapitre **D** s'intéresse au **niveau microscopique** de la gestion documentaire qui est centré sur l'activité de RI et plus précisément sur l'amélioration des systèmes de recherche d'information (recherche d'information par requêtage) en tant qu'outils de recherche les plus utilisés. Ainsi les principales problématiques rencontrées par les utilisateurs des systèmes de recherche d'information sont abordées dans nos travaux : la problématique de la formulation des besoins, la présentation des résultats de recherche d'information et le contrôle de la RI.
    - la difficulté de **formulation d'une requête** renvoie au problème de la requête qui nécessite une implication forte de la part de l'usager même si cela paraît simple,
    - la **présentation des résultats de recherche** renvoie au problème de l'identification, dans la liste de résultats (même réordonnée et filtrée) des documents répondant aux besoins de l'usager surtout lorsque son objectif est de trouver toutes les informations relatives à un thème par exemple. Cet objectif est utopique dès l'instant où le système de recherche retourne une liste de résultats composée de plusieurs milliers de documents comme c'est très fréquemment le cas dans le contexte du Web. Le problème n'est pas lié à la qualité du système qui retourne les documents pertinents (du point de vue système), il s'agit plutôt de s'interroger sur la **présentation des documents** retrouvés par le système à l'usager.
    - le **contrôle de la RI** renvoie au problème des traitements annexes que l'usager doit mettre en place pour exploiter de façon adaptée les très nombreux résultats de recherche. Cette exploitation des résultats a pour objectif de réaliser une tâche précise qui repose sur les informations recherchées.
- Le chapitre **E** s'intéresse à une problématique transversale de la gestion documentaire, c'est-à-dire aux modèles pour la personnalisation. Nous abordons ainsi la problématique des modèles de profils et de contexte. Ces modèles sont appliqués à la RI dans ce mémoire à des fins d'illustration.

C. Amélioration Des Activités Documentaires : Approches Centrées sur la RI

C.1 Introduction

Molinaro et al. dans (Molinaro, 2004) souligne la relation forte entre information et document : « Le couple information/document, qui paraît donc inséparable, remplit le rôle de matière première capitale pour l'entreprise ». En effet, nous avons précédemment souligné que, dans le contexte du Web, la recherche documentaire et la recherche d'information étaient fusionnées. Cette activité de recherche est le plus souvent initiatrice d'un processus de documentation et plus particulièrement de création et de diffusion documentaire. Une étude chiffrée indique qu'il existe une catégorie d'employés, qualifiés de « knowledge worker », qui représentait 31 % de la main d'œuvre des États-Unis en 1995, cette proportion devant continuer à croître de manière significative au XXIème siècle (Sellen, et al., 2003). Ces « travailleurs du savoir » travaillent principalement avec des informations afin d'en produire de nouvelles. Ainsi, la recherche d'information n'est que rarement une fin en soi. Elle s'intègre dans un processus plus large de GED (Gestion Électronique de Documents) ou GEIDE (Gestion Électronique d'Informations et de Documents Existants). La GEIDE est apparue en 1994 sur proposition de l'association APROGED[11] (Association des professionnels pour l'économie numérique) qui souhaitait élargir la notion de GED notamment à tout type d'information. Cependant, malgré la différence d'acronyme les objectifs, techniques et solutions sont sensiblement les mêmes. Dans ce mémoire, nous utiliserons uniquement l'acronyme de GED.

La GED peut être définie comme « un ensemble de techniques permettant d'organiser, de gérer et de distribuer des informations sous forme électronique mais également de structurer des documents pour favoriser l'utilisation, la circulation, les échanges et l'archivage : documents numérisés, forum structuré, architecture, sécurité, aspects juridiques » (Humbert, 2005).

Pour réaliser une GED, de nombreux outils et systèmes, qu'ils soient libres ou propriétaires, ont été proposés sur le marché depuis de nombreuses années. Un comparatif de 10 systèmes réalisé en 2009[12] présente trois types de GED :

- Les GED simples : « Faciles à utiliser et à mettre en œuvre, elles permettent de numériser, classer, partager et archiver les contenus »,

[11] http://www.aproged.org/

[12] http://sites.google.com/site/conseilscomparatifsgedgratuits

- Les GED orientées traitements de masses : « Elles permettent de numériser, classer, partager et archiver de gros volumes de contenus pour de longues durées »,
- Les GED collaboratives : « Outre la gestion documentaire, elles offrent des capacités d'intégration au SI et des fonctionnalités de travail collaboratif (publication les contenus sur des plateformes web 2.0, etc) ».

Ces différents types de GED diffèrent notamment par le coût d'acquisition, la complexité d'installation ainsi que par les fonctionnalités intégrées. Cependant, même si les fonctionnalités sont différentes d'un système à un autre, nous pouvons souligner que tous les systèmes reposent sur le cycle de vie du document au travers duquel sont identifiées les principales **activités documentaires**. Nous pouvons souligner l'importance de l'activité de RI dans ce cadre.

C.1.1 Les activités documentaires

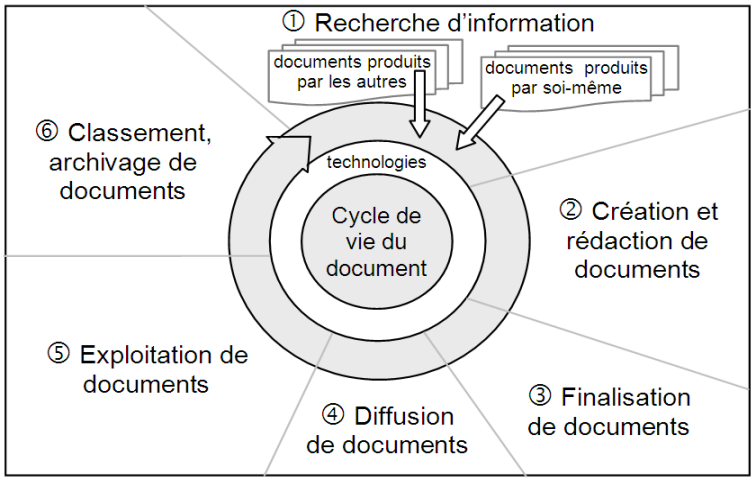


Figure 8 - Cycle de vie d'un document (Sellen, et al., 2003)

Six activités documentaires (Figure 8) liées au cycle de vie d'un document ont été identifiées (Sellen, et al., 2003). Nous les présentons ici en mettant en évidence les inconvénients qui leur incombent :

- La Recherche d'Information ①. Cette activité est très chronophage pour les usagers. Selon une étude de Feldman (Feldman, 2004) entre 15 % et 35 % du temps de travail des individus dans un contexte organisationnel concerne la RI. Il souligne également que cette activité est peu rentable car la moitié des recherches échoue et que dans un contexte professionnel près de 40% des sujets de l'étude ne trouve pas l'information utile pour leur activité au sein des sources internes (intranet…). La Figure 8 souligne le fait que la RI débute sur la base d'informations que l'utilisateur trouve soit à l'extérieur (moteur de recherche sur le Web par exemple) soit localement dans son propre espace personnel d'information (les signets, les tags, les répertoires qu'il a lui-même créés etc…). Cet espace personnel d'information sera noté par la suite **EPI** et sera développé dans l'activité ⑥,

- Les activités de création, rédaction ② et finalisation ③ de documents sont principalement mises en œuvre grâce à des logiciels de traitement de texte. Il existe également des approches complémentaires comme les wikis qui

permettent la rédaction collaborative asynchrone. D'autres approches rendent la rédaction synchrone possible : chaque rédacteur voit en temps réel les modifications des autres contributeurs (Swarts, 2004). Cette activité possède un faible rendement, en partie dû à l'inefficacité de l'activité de RI ① : Feldman indique que selon différentes études, les utilisateurs passeraient 90% de leur temps à recréer de nouveaux documents à partir d'informations qui existent bel et bien dans l'organisation mais qu'ils n'ont pas pu localiser,

- La diffusion de documents ④. En effet, créer un document n'est pas une fin en soi et il existe principalement deux stratégies pour diffuser les documents. La stratégie manuelle, qui repose notamment sur l'envoi de courriels, l'envoi d'annonces sur des listes de diffusion thématiques, ou la publication sur l'intranet d'une l'organisation. Ces actions représentent une charge cognitive importante pour l'usager : pour chaque document à diffuser, il est confronté à l'identification des destinataires potentiellement intéressés au sein de l'organisation, ainsi qu'au choix de la rubrique adéquate dans l'intranet. Afin de soulager les usagers de tels efforts rédhibitoires, la stratégie automatique comprend les moteurs de workflow, les systèmes de recommandation, ainsi que des approches fondées sur l'élicitation et l'exploitation de réseaux sociaux (Ehang, et al., 2005),

- L'exploitation de documents ⑤ correspond notamment à l'activité de lecture active définie dans (Adler, et al., 1972). Celle-ci permet à l'individu de s'approprier le contenu d'un document grâce à la réflexion critique qu'il mène durant sa lecture. Cette tâche implique de la part du lecteur la création d'annotations informelles telles que des commentaires, des résumés, des marques diverses, etc. Kidd dans (Kidd, 1994) indique que ces marques, qui enrichissent le document afin d'extérioriser sa pensée, seront par la suite utiles pour générer de l'information. Dans (Sellen, et al., 2003) les auteurs rapportent également que la pratique d'annotation est importante pour les travailleurs du savoir, car elle leur permet de structurer et d'organiser leur pensée. Bien qu'elle soit répandue et facile à mettre en œuvre sur papier, c'est une pratique pas ou peu supportée — du moins avec moins de souplesse que sur le papier — dans les environnements informatiques, suscitant de fait la frustration des lecteurs lorsqu'ils sont privés de cet outil précieux,

- L'activité de classement et d'archivage de documents ⑥ permet à l'usager de stocker, notamment dans son EPI, les documents importants et utiles pour son activité Dans (Kaye, et al., 2006) sont soulignées les principales motivations de classement et d'archivage : retrouver des documents plus tard, construire un héritage informationnel (legacy) et les partager. De façon traditionnelle, les usagers recourent généralement au classement hiérarchique comme l'est notre système de fichiers, en les décomposant éventuellement de la même manière que les projets de l'usager sont décomposés en sous-projets (Jones, 2007). Cette stratégie est confirmée dans (Khoo, et al., 2007), qui soulignent la fréquence des arborescences décomposées selon un à trois niveaux au moins. L'activité de classement d'un document dans l'EPI est une tâche hautement cognitive pour l'usager. Rücker et al. dans (Rücker, et al., 1997) indiquent que le regroupement d'objets dans un répertoire reflète une cohérence sémantique entre ces objets. Pour ce faire, l'individu doit tout d'abord se construire une représentation mentale du contenu d'un document puis décider de le conserver ou pas en

estimant son utilité future, par anticipation. Le cas échéant, il doit alors identifier le répertoire le plus adapté dans son EPI, éventuellement en créer un nouveau, pour enfin y stocker le document considéré. Cette activité de classement est fréquemment mise en œuvre dans le contexte du Web : (Abrams, et al., 1998) rapportent notamment qu'un usager de signets Web en crée de trois à quatre par session de navigation. Cette activité a évolué dans sa forme avec les plateformes d'étiquetage social présentées précédemment.

Au regard des systèmes existants, aussi performants qu'ils soient, il apparaît que les différentes fonctionnalités correspondant aux différentes activités documentaires restent **cloisonnées** c'est-à-dire qu'il n'existe pas ou peu de dynamique (relations) entre elles mêmes si elles reposent sur un corpus de documents partagés. Ceci limite ainsi la portée de la personnalisation apportée aux différents systèmes permettant de réaliser les différentes activités. La raison de cet état de fait est que la plupart des systèmes proposent des fonctionnalités qui ne couvrent qu'une voire deux activités documentaires ne possédant ainsi qu'une vision partielle de l'activité globale de l'usager.

Dans le même temps, **l'interopérabilité** entre les systèmes est quasi inexistante. Ceci a pour incidence pour l'usager de devoir toujours « repartir de zéro » dans chaque activité et surtout pour les systèmes utilisés de n'avoir qu'une vision incomplète de l'usager (du point de vue profil) étant donné qu'ils n'ont pas l'intégralité des informations liées à l'usager à leur disposition. Cet aspect est fortement dommageable pour l'usager car son activité est nécessairement ralentie par le fait qu'il doive redonner son « contexte » à chacune des applications qu'il utilise.

Ce problème est également accentué par la nécessité pour l'usager de gérer et de maintenir **plusieurs arborescences** de documents (fichiers, emails, signets, etc.), impliquant une fragmentation des données dont les usagers se plaignent comme indiqué dans la synthèse des recherches sur la mémorisation et l'organisation des informations personnelles (Jones, 2007).

Enfin, **l'aspect collectif** est encore sous-développé dans les activités documentaires (même si cet aspect tend à s'estomper notamment dans les systèmes de GED collaboratifs) ne permettant pas pour chaque usager de profiter des résultats des autres usagers.

Il résulte de cette situation une faible exploitation et pérennisation de l'information introduite dans cette organisation. Pourtant l'usager y gagnerait en performance et en dynamisme si ces activités documentaires étaient « liées » et si elles s'enrichissaient mutuellement.

Dans ce contexte, nous proposons une approche centrée sur une fédération des activités de l'usager permettant l'enrichissement mutuel des activités documentaires afin d'améliorer l'activité globale des usagers.

C.2 Enjeux & Objectifs de nos travaux

Dans ce chapitre, nous présentons une approche originale qui favorise la relation usager/information tout en améliorant l'activité documentaire globale des usagers. Cette approche repose sur :

- un **enrichissement mutuel** des différentes activités documentaires pour un usager pour améliorer l'efficacité globale de l'usager dans sa gestion

documentaire. Cet enrichissement repose sur une fédération des activités documentaires favorisant l'échange entre les différentes activités documentaires,

- une approche centrée sur l'usager et essentiellement basée sur des processus **adaptatifs** favorisant le partage d'information entre les différents usagers,
- une **dimension collective**, c'est-à-dire que chacun participe au travers de ses activités à l'amélioration des activités documentaires des autres usagers,
- un respect de la vie privée.

Un des points forts de l'approche proposée est qu'elle reste proche de l'activité réelle des usagers car elle repose sur une pratique courante et développée dans le cadre de la GED : la pratique d'annotation. Notre approche est également centrée sur l'activité commune de recherche d'information.

L'enrichissement mutuel que nous proposons consiste à fédérer les différentes activités documentaires autour du concept **d'annotations collectives** qui sert alors de vecteur d'intégration. Cette activité d'annotation est séculaire et est encore très utilisée par les usagers du fait qu'elle ait été portée du support papier aux supports informatiques grâce aux systèmes d'annotation.

C.3 Définition du concept d'annotation collective

C.3.1 Contexte des Annotations

Une annotation peut être définie comme suit (Bringay, et al., 2004) :

« Une **annotation** est une note particulière attachée à une **cible**. La cible peut être une collection de documents, un document, un segment de document (paragraphe, groupe de mots, mot, image ou partie d'image, etc.), une autre annotation. À une annotation correspond un contenu, matérialisé par une inscription, qui est une trace de la représentation mentale que l'annotateur se fait de la cible. [...] Nous appelons l'**ancre** ce qui lie l'annotation à la cible (un trait, un passage entouré, etc.) ».

Dans cette définition, nous pouvons distinguer ce que nous appelons dans ce mémoire une annotation, que nous qualifierons d'informelle, c'est-à-dire créée par un usager. Elle diffère de l'annotation automatique ajoutée par un système. Par ailleurs, nous pouvons souligner le concept d'ancre qui peut, comme souligné dans la définition, revêtir différentes formes souvent associées à des objectifs différents :

- Forme **textuelle** : cette forme basée sur du texte libre permet à l'usager d'ajouter des notes de lecture, des remarques, des corrections… Cette forme d'annotation est majoritairement indiquée dans la marge des documents (Ovsiannikov, et al., 1999),
- Forme **non textuelle** : cette forme d'annotation repose sur des symboles (choisis librement par l'usage) correspondant à des objectifs dont les principaux sont :
    - La mise en valeur d'un passage du document. La forme la plus courante est le surlignement ou le soulignement (Ovsiannikov, et al., 1999), (Denoue, 2000).

- le typage et la catégorisation des passages du document annoté. Dans cet objectif les couleurs peuvent jouer un rôle important permettant de rapprocher visuellement les passages ou les annotations.
- re-segmenter le document. L'usager indique par le biais d'annotations une structuration du document qui lui est propre grâce à des flèches notamment.

Ces différentes formes d'annotation peuvent cohabiter dans le même document comme illustré dans la Figure 9.

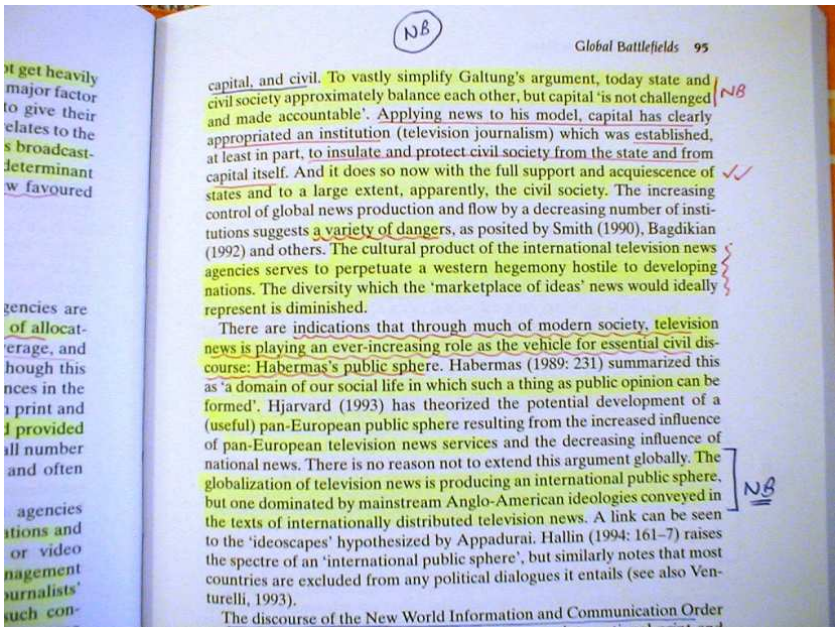


Figure 9 - Page d'un livre annoté de la bibliothèque universitaire de Cambridge

En terme de finalités pour l'usager, l'annotation joue un rôle important, notamment grâce aux différentes formes qu'elle revêt, pour :

- **se remémorer les points clés d'un document**,
- **favoriser la lecture active**. Cette dernière consiste à s'approprier le contenu d'un document tout en menant une analyse critique de son contenu. Ce type de lecture a pour objectif d'extraire du document les éléments importants pour pouvoir par exemple les exploiter (activité documentaire n°⑤ du cycle de vie du document c'est-à-dire l'exploitation des documents), pour créer un document par exemple. L'annotation est une pratique indispensable pour mener à bien cette lecture active (Adler, et al., 1972), (Marshall, 1998), (Jackson, 2002),
- **catégoriser des passages de documents**,
- **matérialiser physiquement l'état d'avancement d'une tâche**.

Sur un plan collectif, les annotations (principalement textuelles) sont considérées comme une valeur ajoutée au contenu des documents (Wolfe, 2000) dès l'instant où les documents sont partagés. Elles revêtent alors un rôle de communication et d'échanges entre les différents lecteurs sur le contenu des documents. Ainsi elles permettent à leurs auteurs d'ajouter des opinions ainsi que des critiques sur le contenu des documents. Ces ajouts permettent aux futurs lecteurs de mieux apprécier le contenu du document qui au départ ne relatait que le point de vue de l'auteur. La lecture active est du même coup facilitée.

C.3.2 Concept d'annotation collective

Dans ce contexte et sur la base d'une revue de littérature autour des systèmes d'annotations développés entre 1989 et 2008, nous avons défini le concept **d'annotation collective** (Cabanac, et al., 2005). Ces annotations sont représentées par le couple <*DO*, *IS*>.

Où :

- *DO* représente l'ensemble des données objectives de l'annotation telles que :
  - **l'identité** de son auteur,
  - une **estampille temporelle** permettant de savoir à quel moment cette annotation a été créée,
  - les **points d'ancrage** correspondant aux parties du document auxquelles se rapporte l'annotation,
  - un **fil de discussion**.
- *IS* représente l'ensemble des informations subjectives optionnelles qui sont fournies par le créateur de l'annotation telles que :
  - un **contenu** permettant au créateur de l'annotation d'ajouter un commentaire textuel,
  - une **visibilité** permettant au créateur de choisir s'il souhaite donner ou non accès à l'annotation aux autres usagers,
  - **l'expertise** de son créateur dans le domaine abordé par le passage annoté. Cette expertise est fixée subjectivement par le créateur de l'annotation. (Marshall, 1998) indique que cette indication est utile pour les futurs lecteurs qui tendent à faire davantage confiance aux experts qu'aux débutants,
  - un **jugement de pertinence** explicite quant au passage annoté (exemples : j'aime, indifférent, je n'aime pas),
  - une liste de **références** fournie par le créateur de l'annotation permettant d'établir des liens entre le passage annoté, le contenu de l'annotation et des ressources externes. Ce type d'informations accroît encore les possibilités de description des annotations,
  - une liste de **tags** permettant de décrire l'annotation. L'objectif est le même que dans les plateformes d'étiquetage social (Hammond, et al., 2005), c'est-à-dire le classement dans un EPI et le partage des annotations,
  - une liste de **types** d'annotation permettant au créateur de l'annotation de fournir un aperçu de la sémantique de son annotation. Les types proposés dans notre approche sont basés sur les types proposés par (Kahan, et al., 2002), auxquels nous avons adjoint des types d'opinion afin de répondre aux besoins des lecteurs identifiés dans (Wolfe, 2000). Nous avons divisé les types d'annotation en deux classes : « commentaire » et « opinion » (Tableau 2). Le type correspondant à l'opinion traduit la position du créateur de l'annotation quant au passage annoté (exemples : d'accord ou pas, neutre). Une annotation peut être qualifiée par aucun, un ou plusieurs types de « commentaire » synthétisant ainsi son contenu.

Classe	Commentaire			Opinion (types exclusifs)		
Type	question	modification	exemple	réfutation	neutre	confirmation
Notation	$\mathscr{Q}$	$\mathscr{M}$	$\mathscr{E}$	$\mathscr{R}$	$\mathscr{N}$	$\mathscr{C}$

Tableau 2 - Types d'annotation sur lesquels repose notre approche

La variété des informations associées à une annotation permet de couvrir les différentes finalités d'une annotation qui seront utiles pour la fédération des activités documentaires. Nous nous sommes limités dans notre approche aux trois finalités les plus importantes de notre point de vue qui sont : **mémoriser, commenter et débattre.** Selon chaque type, tout ou partie des informations de l'annotation pourront être renseignées.

Les annotations ayant pour finalité de « **mémoriser** » représentent une extension des signets Web classiques très utilisés par les usagers. L'annotation, au travers de son point d'ancrage, permet de référencer tout ou partie du document annoté contrairement au signet qui fait référence au document complet. Ces annotations représentent donc le point central de l'activité n°⑥ (classement) du cycle de vie du document. Ces annotations seront stockées dans un EPI que nous avons étendu en EPA (Espace Personnel d'Annotations) pour supporter les différents types d'annotations proposées. Cet EPA comme les EPI supporte la classification hiérarchique des informations.

Les annotations ayant pour finalité de « **commenter** » permettent aux lecteurs de signaler des remarques de forme, des coquilles… elles sont liées essentiellement à l'activité documentaire n°③ (finalisation de documents). La différence avec les autres annotations est qu'elles ne nécessitent pas nécessairement un stockage dans l'EPA de leur créateur.

Les annotations ayant pour finalité de« **débattre** », que nous nommons également *annotations collectives argumentatives*, permettant au créateur de faire des critiques sur le fond du passage annoté. Une discussion pourra alors être suscitée, les arguments pourront ainsi être échangés et partagés entre les lecteurs au travers du fil de discussions associé àl'annotation. Dans notre approche, chaque réponse à une annotation est considérée elle-même comme une annotation (possédant les mêmes caractéristiques).

En résumé, les annotations collectives que nous proposons d'utiliser (Cabanac, et al., 2005) :

- **intègrent la plupart des propositions de la littérature** (exemple : types d'annotations),
- permettent de **couvrir les différentes finalités** attendues par les annotations notamment grâce à leur richesse de contenu,
- **favorisent les échanges** entre lecteurs avec le support notamment du fil de discussion.

Les annotations ainsi définies peuvent être appliquées à différents formats de fichiers (textuels, images, vectoriels..) au point d'ancrage près. Cependant, dans ce mémoire, nous nous limitons aux seuls documents textuels et particulièrement aux documents textuels du Web car il est aujourd'hui une source d'information privilégiée pour réaliser les activités documentaires.

C.4 Usage des annotations pour la fédération des activités documentaires

La pratique d'annotation se retrouve « nativement » dans certaines activités documentaires : n°②, ③ et ⑤ (respectivement création, finalisation et exploitation de documents). Par ailleurs, au regard de la définition de l'annotation collective que nous proposons, notamment au travers de la finalité de mémorisation, l'activité de classement et d'archivage ⑥ peut donc être assurée naturellement au travers des annotations.

Pour les autres activités (n° ① et ④) correspondant respectivement à la RI et à la diffusion d'information, (Fraenkel, et al., 1999) souligne l'intérêt de l'utilisation des annotations pour améliorer le processus de recherche d'information.

Outre le fait que les annotations jouent un rôle important dans différentes activités documentaires, l'intérêt que nous avons trouvé aux annotations est qu'elles sont **contextuelles** car chacune d'entre elles est relative à un passage spécifique du document annoté. La discussion qui peut en découler est donc elle-même contextuelle.

De ce fait, étant donné que l'annotation peut être exploitée dans les différentes activités documentaires, elle peut jouer le **rôle de vecteur d'intégration de ces activités**.

C.5 Fédération des activités documentaires par l'annotation collective

Pour atteindre notre objectif de fédération des activités documentaires, nous avons étudié le cycle de vie du document afin d'identifier les « ponts » qui pouvaient être établis entre les différentes activités. A la suite de cette étude, nous avons identifié les processus permettant d'améliorer l'activité de l'usager sur la base des annotations collectives organisées au travers des EPA des usagers (Chevalier, 2002), (Chevalier, et al., 2003), (Cabanac, et al., 2006d), (Cabanac, et al., 2007). Ces EPA sont en effet une mine d'informations à forte valeur ajoutée car leur construction est le résultat d'un effort cognitif important de la part des usagers (Rücker, et al., 1997).

En résultat, sur la base des travaux réalisés, nous avons identifié cinq processus originaux visant à exploiter ces mines d'informations afin d'améliorer les différentes activités selon un principe d'enrichissement mutuel. Cela signifie que, par exemple, lors d'une navigation (activité n°①), les documents visités pourront être diffusés aux autres usagers pour qu'ils puissent en prendre connaissance (proposés par exemple dans leur EPA – activité n°①). Dans le même temps, le fait que les usagers classent et maintiennent (activité n°⑥) des références vers des documents (ou parties de documents) permet au système de découvrir des liens implicites entre ces documents qu'il peut exploiter pour, par exemple, recommander des documents au cours de la navigation d'un usager (activité n°①).

La Figure 10 (Cabanac, et al., 2008), (Cabanac, 2008b), (Cabanac, et al., 2009) présente les différents processus qui, sur la base des EPA, participent à l'enrichissement mutuel des activités. Dans cette figure, deux usagers sont représentés. Nous soulignons le fait que la plupart des processus sont dédiés à un usager unique (indicé par l'usager) à l'exception du processus *Reco* qui est partagé par l'ensemble des utilisateurs.

Nous présentons dans les sections suivantes les différents processus proposés.

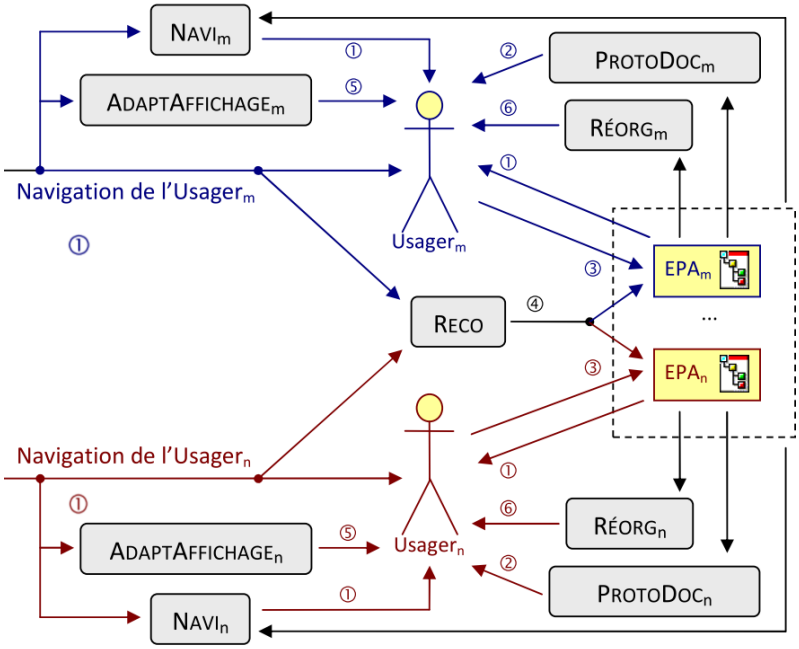


Figure 10 - Architecture générale proposée pour la fédération des activités documentaires

C.5.1 Processus *Navi*

Ce processus adaptatif et individuel exploite l'historique de navigation d'un usager ainsi qu'une source d'évidences pour recommander des documents au cours de la navigation (activité n°①). Une source d'évidences correspond à une source d'informations permettant au système d'extraire des indicateurs relatifs, dans notre cas, à la pertinence des documents pour la navigation des usagers. Ces sources peuvent être des outils externes tels que les outils de recherche, les historiques de navigation…

L'approche que nous avons proposée (Chevalier, et al., 2002b), (Chevalier, et al., 2003), (Chevalier, et al., 2004) repose sur un profil usager constitué d'un **historique de navigation** associé à une liste de recommandations sur laquelle une fonction **d'oubli graduel** est appliquée. La « **traçabilité** » des documents recommandés et présentés à l'usager a été privilégiée. Ainsi, l'usager peut comprendre pourquoi un document lui a été recommandé et d'où il provient.

L'originalité de notre approche consiste à utiliser, pour identifier les documents à recommander, une source d'évidences originale et sous-exploitée : les EPA de l'ensemble des usagers. Au travers de ces EPA nous exploitons une mesure de similarité : **la similarité d'organisation** (ou d'usage). Cette similarité permet d'identifier des liens implicites entre les documents et ceux qui ont été visités par l'usager. Elle est basée sur la façon dont ils sont organisés dans les EPA et non sur leur contenu. Cette similarité d'organisation est présentée dans la section C.6.1 p. 45. Un autre point notable est l'attention apportée à la relation entre l'usager et la traçabilité des informations.

Une évaluation de ce processus est présentée dans (Chevalier, 2002). Elle souligne que le processus proposé permet d'obtenir au fil de la navigation un ensemble de documents recommandés dans lequel les documents pertinents pour la navigation globale de l'usager sont situés dans les premières recommandations alors que les documents non pertinents se trouvent en fin de liste.

Ce processus est assez proche des systèmes de recommandation proposés dans la littérature tels que *Letizia* (Lieberman, 1995), *WebWatcher* (Armstrong, et al., 1995) ou

encore *Broadway* (Trousse, et al., 1999). Ces systèmes recommandent en effet des documents liés par des liens hypertextes aux documents visités par l'usager. Notre approche est plus générale et propose des documents non nécessairement liés par des liens hypertextes aux documents visités. Ainsi, notre proposition se rapproche de celle de *WBI* (Barrett, et al., 1997) qui repose sur un outil de recherche pour trouver et recommander des documents proches thématiquement des documents visités par un usager. Notre approche se démarque par le fait qu'elle ne recommande pas de documents similaires en fonction de leur contenu mais en fonction de leur organisation dans les EPA des usagers. Elle se dénote également par l'usage d'un historique de navigation ainsi qu'une fonction d'oubli graduel qui permet de recommander des documents pour la navigation complète de l'usager et non uniquement pour le dernier document visité.

C.5.2 Processus *ProtoDoc*

Ce processus individuel permet à l'usager de construire, sur la base d'un ensemble d'annotations, un *proto-document* (activité ②). Un proto-document peut être vu comme une collection de matériels documentaires, plus ou moins cohérents et organisés (Pédauque, 2006).

Dans une première approche, le principe proposé repose sur l'intégration dans le proto-document des passages annotés auxquels les annotations sélectionnées sont rattachées. Ces informations peuvent être organisées dans le proto-document sur la base de leur contenu (classées par similarité de contenu par exemple).

Ce processus repose donc sur l'activité d'annotation de façon générale (notamment réalisée au cours des activités ⑤ et ⑥). Il permet de générer un document à partir du contenu des annotations et des textes annotés.

C.5.3 Processus *Reco*

La diffusion d'information n'est pas une activité évidente surtout du point de vue de l'usager. Feldman dans (Feldman, 2004) souligne à quel point la diffusion d'information est difficile et limitée au sein d'une organisation. En effet malgré une multitude d'outils disponibles (workflow, email, flux rss…) identifier les personnes susceptibles d'être intéressées et à qui envoyer une information n'est pas chose évidente car cela nécessite une connaissance importante tant des personnes, que de leurs centres d'intérêts, de leurs besoins, etc.

En réponse à cette difficulté, le processus adaptatif et collectif *Reco* (Chevalier, et al., 2003b), (Chevalier, et al., 2004) repose sur une approche de recommandation. Sur la base des documents introduits dans l'organisation (par le biais de la RI notamment) ce processus recommande de nouveaux documents directement dans l'EPA des usagers qui le souhaitent.

L'originalité de notre approche est qu'elle permet à toute personne de profiter de l'activité de recherche des autres membres du groupe. L'approche proposée repose sur une approche basée sur un profil usager composé de **classifieurs**. Ainsi, chaque répertoire de l'EPA d'un usager est associé à un classifieur qui indiquera par la suite si un nouveau document est pertinent ou non pour le répertoire associé. Dès qu'un document est jugé pertinent pour un répertoire, il est intégré dans l'EPA de l'usager correspondant qui est alors alerté. Un autre aspect notable de notre approche réside dans le fait que les classifieurs sont dépendants les uns des autres et sont reliés **hiérarchiquement** pour tenir compte de l'arborescence éventuelle des EPA. Par ailleurs ces classifieurs utilisent un **seuil de similarité optimal et dynamique** permettant de réajuster au cours du temps leur fonctionnement. Une

attention particulière a été portée à la valeur de ce seuil car il permet au classifieur de faire la distinction entre documents pertinents et documents non pertinents.

Une évaluation de ce processus est détaillée dans (Chevalier, 2002) sur la collection des documents *OHSUMED* utilisée dans la plateforme d'évaluation *TREC* (Voorhees, 2001). Ces documents sont manuellement associés aux répertoires d'une hiérarchie, la *Medical Subject Heading (MeSH)* de la *National Library of Medecine*. La sous-collection utilisée est celle des maladies cardio-vasculaires. Une présentation complète de l'arborescence utilisée dans cette expérimentation est faite en Annexe I (p. 133). L'expérimentation souligne, sur la base d'une comparaison avec les démarches classiques de la littérature, le fait que les classifieurs proposés et exploités de façon hiérarchique permettent d'améliorer sensiblement la pertinence des recommandations réalisées. Ainsi par exemple, nous obtenons en moyenne une qualité de +5,22 % des recommandations grâce à l'organisation hiérarchique des classifieurs proposée par rapport à des classifieurs indépendants. Une autre incidence de la hiérarchisation des classifieurs se retrouve sur le temps de calcul qui peut être réduit de près de 70% par un parcours ascendant plutôt que descendant de la hiérarchie de classifieurs.

Cette approche est assez similaire à l'approche proposée dans (Rücker, et al., 1997) qui recommande aux usagers dans leur EPI des informations détenues dans les EPI des autres usagers. Cependant, l'aspect hiérarchique de la construction des EPI n'est pas pris en compte dans cette approche. Notre approche se démarque toutefois de cette dernière en ce sens qu'elle prend en compte l'aspect hiérarchique permettant d'obtenir de meilleurs résultats.

C.5.4 Processus *Réorg*

Ce processus individuel a pour objectif d'aider l'usager à réorganiser son EPA. En effet, l'usager ajoute successivement des annotations au cours de la lecture de différents documents par exemple. Cependant (Abrams, et al., 1998) et (Jones, 2007) soulignent le fait que l'organisation des annotations dans l'EPA qui est une activité hautement cognitive et qu'elle est souvent reportée dans le temps. On peut constater qu'à force d'insérer des documents dans un répertoire, le besoin de subdivision en sous-répertoires se fait sentir.

Dans nos travaux (Cabanac, 2002), (Chevalier, 2002), nous avons donc proposé une démarche de **Classification Hiérarchique Ascendante** ou CHA (Jardine, et al., 1971) fournissant une arborescence thématique binaire de l'EPA. Nous avons associé à cette CHA une fonction de **seuillage** pour obtenir un nombre de classe optimal, comme proposée dans (Maarek, et al., 1996), ainsi qu'un traitement **d'étiquetage** de chaque classe à partir du coefficient du χ^2. Sur la base du résultat de la classification, l'usager peut accepter toute ou partie de la nouvelle classification pour la reporter sur son EPA. L'usager garde ainsi le contrôle total de son EPA.

C.5.5 Processus *AdaptAffichage*

Ce processus individuel est directement lié à l'usage des annotations. En effet, du fait que les annotations sont le plus souvent intégrées au document, une quantité importante d'annotations (amplifiée par le fait qu'elles soient partagées) peut rendre le texte plus difficilement lisible.

Dans l'exemple présenté dans la Figure 11, les différentes annotations sont signalées dans le texte par des symboles spécifiques (, ,…). Pour voir le contenu d'une annotation, l'usager doit cliquer sur chaque symbole. Ainsi le fait que les symboles utilisés ne révèlent

pas visuellement le contenu complet de l'annotation implique qu'il est difficile pour un usager de faire le tri entre les différentes annotations.

En réponse à cela, nous proposons (Cabanac, 2008) que le symbole de chaque annotation traduise mieux la sémantique de l'annotation. Par ailleurs, le processus ***adaptAffichage*** offre la possibilité de **filtrer** et de **mettre en exergue** certains types d'annotations pour, par exemple, que l'usager puisse se concentrer sur un sous-ensemble des annotations disponibles dans le document. Les critères de filtrage correspondent à l'ensemble des critères décrivant une annotation (section C.3.2). Ce filtrage peut être également réalisé sur la base de l'indicateur de validation sociale que nous avons proposé. Cet indicateur traduit, sur la base du fil de discussion associé le consensus qui existe relativement à l'annotation. Cet indicateur basé sur l'activité d'échanges entre les usagers est décrit dans la section C.6.2.

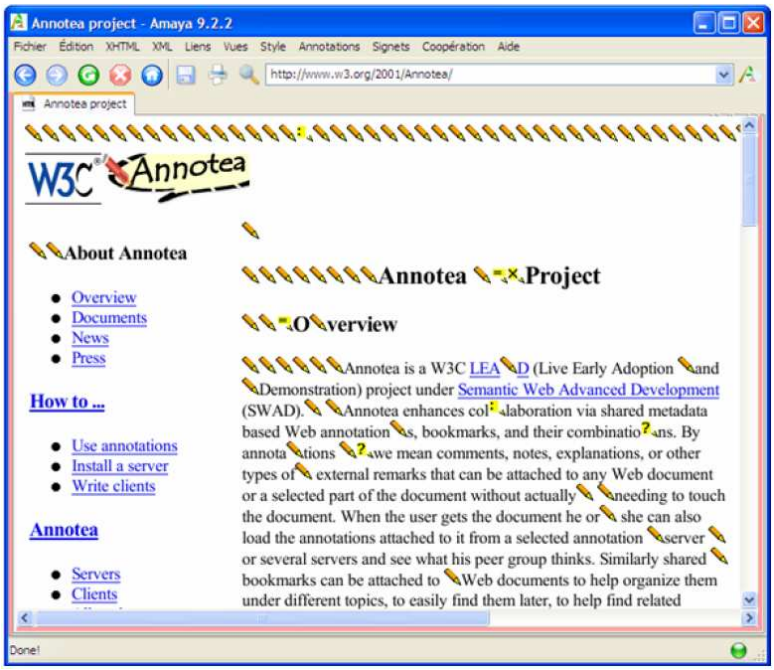


Figure 11 - Impact du nombre d'annotations sur la lisibilité

C.6 Définition d'indicateurs relatifs aux activités documentaires

Les différents processus proposés permettent un enrichissement mutuel des différentes activités documentaires des usagers. Certains processus (*navi*, *adaptAffichage*) reposent sur des indicateurs (Similarité d'Organisation et mesure de validation sociale d'une annotation) que nous proposons. Ces deux indicateurs participent à l'originalité des processus proposés car ils sont centrés sur l'activité des usagers du système.

C.6.1 Similarité d'Organisation (ou d'usage)

La mesure de similarité la plus utilisée en RI est probablement celle basée sur le contenu des documents. Calculer une telle similarité met en jeu les différents concepts (indexation, similarité…) vus dans les sections précédentes.

Nous proposons une mesure de similarité originale : la similarité d'organisation. Celle-ci est directement liée aux EPA et à la connaissance qu'ils contiennent. En effet, les EPA du fait

de leur organisation hiérarchique matérialisent l'effort cognitif important (organisation des répertoires, placer un document dans un répertoire particulier…) comme le souligne notamment (Rücker, et al., 1997). L'objectif de cette similarité est donc d'identifier des relations entre les documents organisés et mémorisés par l'ensemble des EPA des usagers. Elle se base sur une proximité organisationnelle. L'intérêt de cette similarité est qu'elle est complémentaire à la similarité basée sur le contenu.

L'approche proposée repose sur la notion de **multi-arbres** (Furnas, et al., 1994) qui permet de « fusionner » les EPA des différents usagers sur la base des documents qu'ils contiennent (Figure 12).

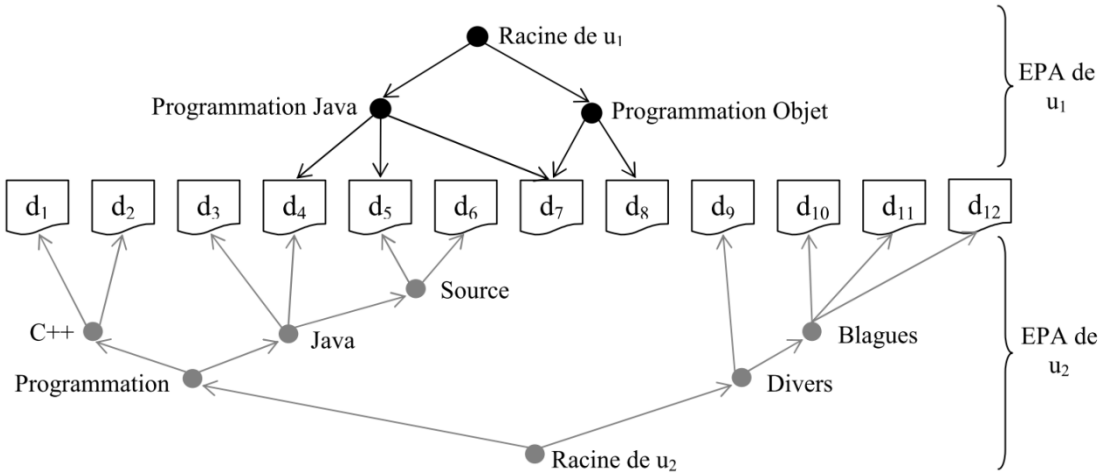


Figure 12 - Exemple de multi-arbres pour deux usagers u_1 et u_2 (d_x correspond à un document)

La mesure de similarité d'organisation que nous proposons (Cabanac, et al., 2007b), (Cabanac, 2008) repose sur les principes suivants :

- (P1) plus la distance moyenne entre deux documents d_1 et d_2 dans les EPA est faible plus leur similarité augmente. La distance utilisée correspond au nombre de sauts entre le répertoire contenant d_1 et celui contenant d_2,
- (P2) plus la proportion d'usagers possédant d1 et d2 dans une même branche de l'EPA est importante, plus le poids social de la similarité est important.

Sur la base de ces deux principes, la distance de similarité d'organisation a été déclinée sur trois types d'objets :

- Entre deux **répertoires** d'un EPA : cette mesure exploite principalement (P1),
- Entre deux **documents** : cette mesure combine (P1) et (P2) et repose sur la similarité entre les répertoires contenant chaque document,
- Entre deux **usagers** : cette mesure utilise la similarité moyenne entre les documents possédés par les deux usagers pour identifier s'ils possèdent et organisent d'une manière proche voire similaire les documents.

Cette fonction de similarité d'organisation permet d'extraire une part de la connaissance contenue dans les EPA. Elle formalise des liens implicites qui existent entre les documents, les répertoires et les usagers.

Pour valider cette mesure de similarité, nous avons réalisé une expérimentation visant à vérifier que cette similarité était bien différente de la similarité de contenu.

Une première validation empirique a permis de constater la différence entre ces deux nous pouvons constater la différence entre ces deux mesures de similarité pour un exemple d'EPA « jouet » (Figure 13).

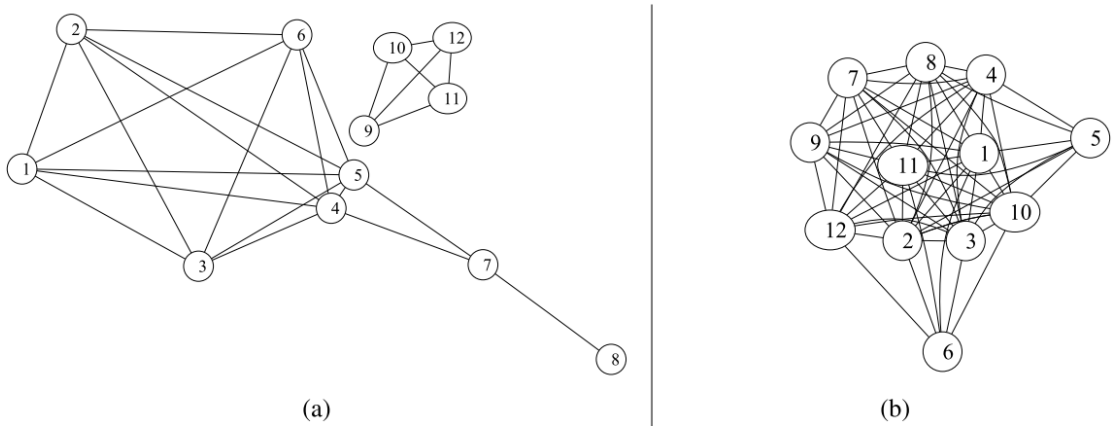


Figure 13 - Différence visuelle entre similarité d'organisation (a) et similarité de contenu (b). Documents correspondant au multi-arbres présenté dans la Figure 12.

Pour vérifier l'apport de la mesure de similarité d'organisation et pour souligner son aspect complémentaire par rapport à la similarité de contenu, une expérimentation plus poussée a été menée. Pour cela nous avons utilisé la sous-collection des maladies cardio-vasculaires *OHSUMED* présentée ultérieurement.

Pour vérifier la différence entre la similarité de contenu et la similarité d'organisation proposée, nous avons comparé les résultats des deux mesures (similarité entre deux documents basée sur le contenu (approche classique) et basée sur l'organisation) en utilisant les tests statistiques d'hypothèse (Savy, 2006). Nous avons utilisé pour cela les tests de Student (*t-test*) et de Wilcoxon sur échantillons appariés (Wilcoxon, 1945), fournissant une valeur de significativité p permettant d'interpréter l'issue du test. Une valeur $p < \alpha = 0{,}05$ (Hull, 1993) traduit une différence statistique significative entre les deux séries examinées. De plus, cette différence est d'autant plus avérée que $p \to 0$.

Par ailleurs, nous avons évalué le coefficient de Pearson $r \in [-1;1]$ qui permet d'évaluer le degré de corrélation entre les deux mesures de similarité. Ces méthodes sont d'autant plus proches que $r \to 1$. Une corrélation inverse (c'est-à-dire $x = y^{-1}$) est mise en évidence par $r \to -1$. Enfin, $r \to 0$ est révélateur de deux séries non corrélées.

Indicateur	Valeur
Tests de significativité (Student, Wilcoxon) p	p = 0.000 **→ différence statistiquement significative**
Coefficient de Pearson r	r = 0.154 **→ absence de corrélation**

Tableau 3- Résultat de nos expérimentations permettant de souligner la différence entre la similarité basée sur le contenu et la similarité basée sur l'organisation sur l'exemple de multi-arbres présenté en Figure 12

Ces résultats prouvent que la similarité d'organisation et la similarité de contenu, sur la sous-collection MeSH utilisée, sont différentes et peuvent donc être considérées comme complémentaires dans les applications. Ceci renforce l'originalité des processus reposant sur cet indicateur.

L'originalité de notre approche réside dans le fait qu'elle :

- repose sur le résultat d'une forte activité cognitive : l'organisation des documents → **haute valeur ajoutée**,
- est **généralisable** à tous types de documents (wav, pdf, mov…) puisqu'elle est uniquement basée sur des références vers les documents et non sur le contenu réel des documents,
- est **différente** et considérée comme **complémentaire** de la mesure de similarité de contenu.

Dans le même esprit, nous avons défini un deuxième indicateur basé sur l'activité des usagers. Ce dernier vise à exploiter les débats que les annotations suscitent : la **validation sociale** d'une annotation.

C.6.2 Mesure de Validation Sociale d'Annotations

Certaines annotations que nous proposons peuvent être argumentatives c'est-à-dire qu'elles peuvent être associées à un fil de discussion. Ce fil de discussion permet aux différents lecteurs de donner leur point de vue en contexte (au regard du contenu de l'annotation). La Figure 14 présente un exemple d'annotation associée à un fil de discussion. Une des difficultés pour l'usager face à ce fil de discussion est de se faire une idée sur le contenu de l'annotation initiale « I totally agree ». Pour cela il doit lire les différents arguments des réponses successives pour faire une synthèse du fil de discussion et statuer sur l'argument de l'annotation.

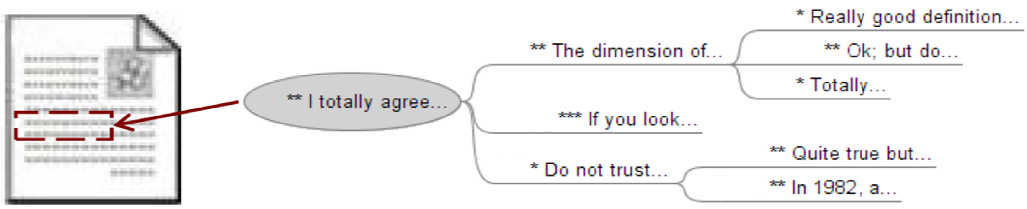


Figure 14 - Un exemple d'annotation avec un fil de discussion associé

L'objectif de cette mesure de validation sociale est de mesurer le consensus autour d'une annotation sur la base du fil de discussion et des caractéristiques des annotations. Il est à souligner que cette mesure repose sur une analyse statistique. Elle ne tient pas compte des arguments réels exprimés comme certaines approches plutôt orientées TALN (Traitement Automatique du Langage Naturel) tels que (Turney, 2002), (Harb, et al., 2008).

La mesure de validation sociale d'une annotation repose sur les principes suivants :

- (P1) plus **l'implication** d'un usager dans une annotation est importante, plus la qualité de l'annotation est importante. Ainsi, une annotation ne contenant qu'un commentaire est considérée de moindre qualité qu'une annotation possédant des xplications, des références étayant le discours ainsi que des exemples. La force de l'opinion (confirmation, réfutation ou neutre) et donc de l'argument de l'annotation est donc corrélée à implication,

- (P2) plus **l'expertise** d'un usager est importante, plus l'annotation est importante,
- (P3) l'argument d'une annotation est modifié par la **synthèse** des arguments des réponses que l'annotation possède (de façon hiérarchique). Si l'argument d'une annotation réfute le texte du document et que tout le monde réfute cet argument alors cet argument sera jugé comme réfuté socialement et donc sans fondement.

Du point de vue définition, la validation sociale est une fonction définie par validationSociale : A → [-1 ; 1] où A est l'ensemble des annotations du système. La valeur de -1 correspond à une réfutation sociale et 1 à une confirmation sociale (0 étant la position sociale neutre). Du fait de cette définition, cette valeur de validation sociale pourra être facilement exploitée par les différents processus informatisés.

Du point de vue du modèle, nous avons exploré trois pistes d'implantation (Cabanac, et al., 2010) venant d'approches différentes pour n'en retenir que deux dans nos expérimentations car l'une d'entre elles ne permettait pas de prendre en compte certaines contraintes que l'on avait.

C.6.2.1 Approche n°1 : degré d'accord entre lecteurs

Cette approche repose sur le coefficient *kappa* de Cohen (Cohen, 1960), et sa généralisation (Fleiss, 1971), utilisé en médecine pour le diagnostic médical fait par différents praticiens. Ce coefficient mesure le degré d'accord entre n juges (personnes) qui répartissent N objets dans k catégories mutuellement exclusives.

Malgré l'intérêt évident de cette approche, ne permettant la prise en compte de l'arborescence des réponses, nous n'avons pas poursuivi l'exploration de cette approche. Nous avons donc proposé une deuxième approche basée sur des heuristiques.

C.6.2.2 Approche n°2 : agrégation récursive de scores d'arguments

Cette approche repose sur démarche heuristique dans laquelle nous proposons un ensemble de fonctions permettant de respecter les principes édictés.

Cette approche purement empirique s'appuie sur un panel de paramètres qui doivent être initialisés à des valeurs optimales. Cette initialisation peut être sensible au contexte rendant cette tâche compliquée. Nous avons donc cherché à limiter le nombre de paramètres devant être adaptés et nous avons identifié une troisième piste qui repose sur une approche plus formelle issue du domaine de l'intelligence artificielle.

C.6.2.3 Approche n°3 : extension d'un système d'argumentation bipolaire

Dans notre étude, nous proposons une extension du système d'argumentation bipolaire proposé dans (Cayrol, et al., 2005).

Définition. Un Système d'Argumentation BiPolaire noté SABP est représenté par le triplet $\langle A, R_{app}, R_{att} \rangle$ où :

- A est un ensemble d'arguments, par exemple : $A = \{a_1, ..., a_n\}$,
- R_{app} est une relation d'appui sur A^2,
- R_{att} est une relation d'attaque sur A^2.

Ainsi un SABP tient compte de deux types de relations entre deux arguments : l'appui (proche de la notion de confirmation) et l'attaque (proche de la notion de réfutation).

L'évaluation graduelle du SABP est définie de façon à respecter trois principes :

- L'évaluation d'un argument est fonction de l'évaluation de tous ses attaquants directs et de tous ses appuis directs,
- si la qualité de l'appui (resp. de l'attaque) augmente alors la valeur de l'argument ainsi appuyé (resp. attaqué) augmente,
- si on ajoute des appuis (resp. des attaques) alors la qualité de l'appui (resp. de l'attaque) augmente.

Fort de ces principes et au regard de la cohérence avec nos objectifs, nous avons adapté le SABP (Cabanac, et al., 2006) pour qu'il puisse prendre en considération les principes (P1) et (P2) de notre approche (l'implication et l'expertise des usagers notamment).

A des fins de validation, nous avons souhaité évaluer les deux approches proposées (heuristique et SABP adapté) en les comparant avec le SABP original (Cayrol, et al., 2005). L'objectif de cette expérimentation consistait à vérifier si les différentes approches informatiques approximaient la synthèse humaine du consensus lié à un fil de discussion.

Face à l'inexistence de protocole d'évaluation nous avons proposé notre propre cadre expérimental (Cabanac, 2008b). Cette expérimentation s'est étalée sur environ 16 mois en 2007 et 2008. Pour cela nous avons construit un corpus de 13 débats argumentatifs. Il est disponible en ligne au format XML à l'URL « http://www.irit.fr/~Guillaume.Cabanac/expe/corpus/ ».

Sur la base de ce corpus, le protocole défini repose sur deux tâches distinctes et ce pour chaque débat :

- Une tâche de valorisation des caractéristiques de chaque annotation (argument) : l'usager doit sélectionner les types d'annotation en termes d'opinion (confirmation, réfutation, neutre) et de commentaire (exemple, modification, question),
- Une tâche manuelle de synthèse des arguments du fil de discussion, c'est-à-dire la perception humaine (*ph)* du consensus autour de chaque argument. Ainsi, pour chaque argument, les usagers devaient indiquer si l'argument est globalement réfuté, confirmé par les arguments fils qu'il possède. Aucune indication sur l'expertise de l'auteur, ni sur les références ne sont disponibles dans le corpus construit.

Après un appel à participation diffusé sur les listes internationales, 121 volontaires ont réellement débuté l'expérimentation. Une description du panel d'évaluateurs, leurs caractéristiques ainsi qu'un détail de l'expérimentation, est fourni dans (Cabanac, et al., 2010).

Après une phase de suppression d'évaluations erronées voire irrationnelles (un argument est réfuté alors que tous ses arguments fils le confirment), nous avons pu vérifier si les algorithmes étudiés approximent ou non la perception humaine. Ainsi pour chaque évaluation, sur la base des caractéristiques affectées par l'usager, nous comparons la perception humaine du consensus (*pH*) fixée par l'utilisateur avec les valeurs obtenues par les approches proposées et ce pour l'argument racine des débats uniquement. En effet, nous souhaitions vérifier la capacité de chaque approche à synthétiser le fil de discussion dans sa globalité.

La différence brute entre les algorithmes et la perception humaine révèle que les trois méthodes ont un comportement assez similaire. Les tests statistiques d'hypothèses valident le fait qu'un problème de démarrage de la phase d'évaluation existe. Ce problème nous a été signalé par différents participants qui ont eu du mal à bien comprendre comment le consensus peut être mesuré ce qui s'améliore au fur et à mesure des débats évalués. Ainsi les premiers débats sont erronés et la différence entre la perception humaine et les approches automatiques proposées s'estompe dès le 4ème débat évalué. Nous précisons que les débats étaient présentés dans le même ordre à tous les usagers. A ce niveau, l'expérimentation démontre que les trois approches permettent d'obtenir des résultats sensiblement identiques.

Pour compléter cette expérimentation, nous avons évalué la différence entre les valeurs de validation sociale des trois approches et la perception humaine en relâchant certaines contraintes telles que :

- la distance : en augmentant la distance acceptable entre la validation sociale calculée et la perception humaine ;
- la force (vérifier que la perception humaine et la validation sociale calculée se trouvent dans une même zone de valeurs) ;
- le signe de la différence entre la perception humaine et la validation sociale calculée (il faut qu'elles donnent le même consensus).

Sur la base de ces dernières expérimentations, la validation sociales calculée (quelle que soit l'approche) est similaire à la perception humaine dans près de 80% des cas. Une différence significative peut cependant être soulignée concernant le temps de calcul. L'algorithme SABP orignal est beaucoup plus rapide (30% plus rapide que SABP étendu par exemple).

L'originalité de notre approche de validation sociale réside dans le fait :

- qu'elle permet d'approximer dans près de 80% des cas la perception humaine,
- qu'un cadre expérimental complet est associé à cette proposition. Les données liées à ce cadre expérimental est mis en disposition en ligne des autres chercheurs.

Les différentes propositions faites dans le cadre de la fédération des activités documentaires ont été implantées au sein de deux prototypes que décrit la section suivante.

C.7 Prototypes

En termes de prototypes, nous présentons le prototype TafAnnote (Cabanac, 2008) qui reprend et étend le prototype *Easy-DoR* développé en 2002 dans le cadre de ma thèse (Chevalier, 2002). Par rapport à nos propositions, seul le processus *Proto-Doc* n'a pas encore été implanté. Les processus *Réorg et Reco* n'ont pas encore été portés de *Easy-DoR* sur *TafAnnote*.

C.7.1 Présentation générale

TafAnnote est développé sous la forme d'une barre d'outils pour le navigateur *FireFox* (Figure 15) et repose sur une architecture client-serveur. Dans cette barre d'outils on retrouve les différentes finalités des annotations collectives (commenter, mémoriser et

débattre). En fonction de la finalité de l'usager, le système lui demandera de renseigner l'ensemble de caractéristiques de l'annotation qu'il souhaite créer (Figure 16). A tout moment, l'usager peut accéder à son EPA (Figure 17). Pour privilégier le confort d'accès, trois modes d'accès à l'EPA sont proposés : un mode par organisation hiérarchique des annotations, un mode par type d'annotation (mémorisation, débattre…) et un autre par tags. Dans Easy-DoR, les documents recommandés par le processus *Reco* sont directement accessibles par cette fenêtre. Pour distinguer les recommandations de ses propres documents, une signalétique spécifique est utilisée (NEW).

Figure 15 - La barre d'outil de TafAnnote dans le navigateur FireFox

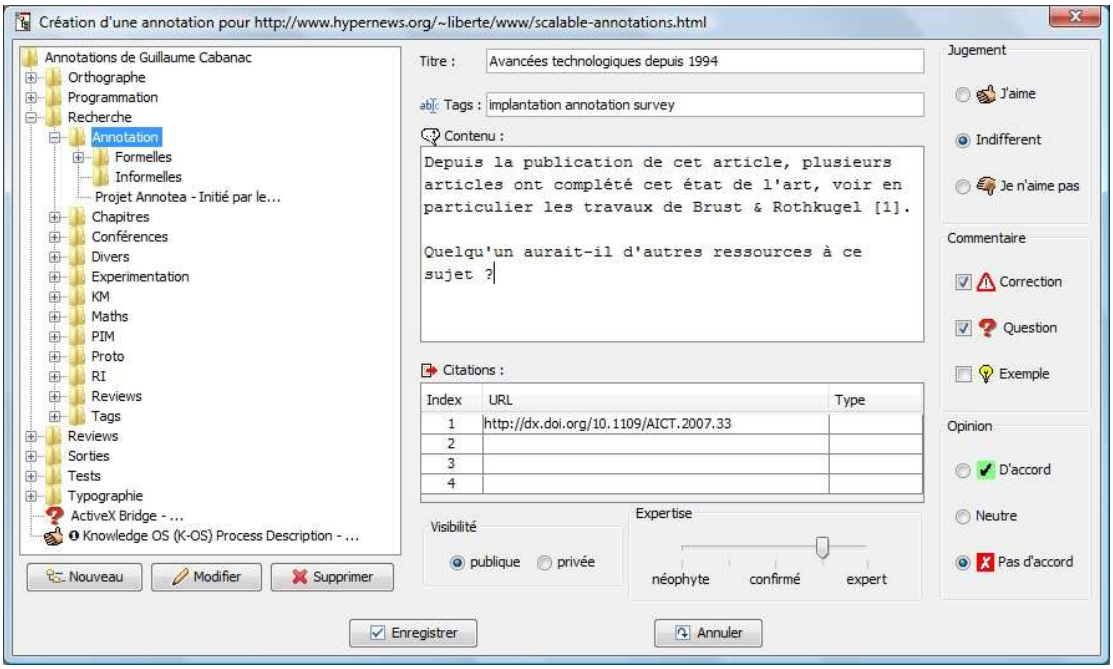


Figure 16 - Création d'une annotation (ici ayant la finalité "débattre")

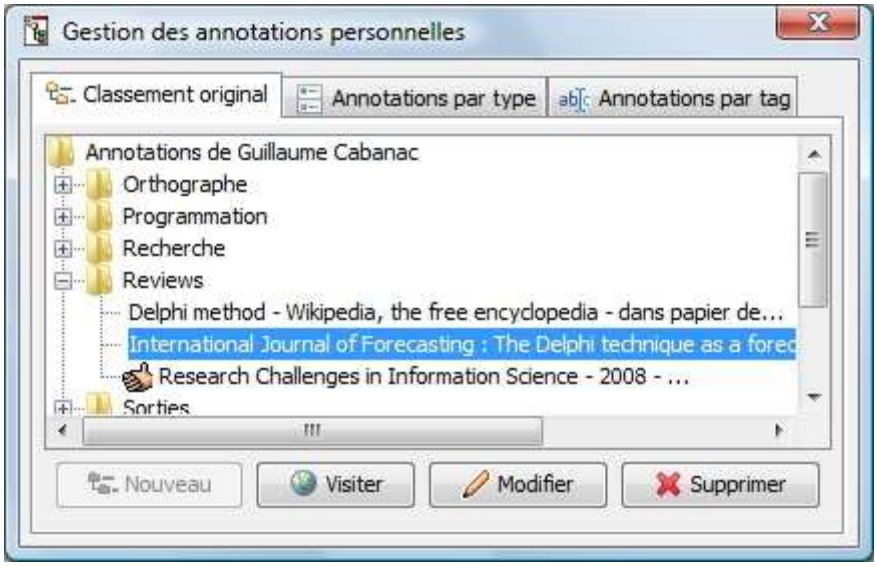


Figure 17 - Fenêtre permettant à l'usager d'explorer son EPA

Au cours de la navigation, et de façon non intrusive, TafAnnote indique à l'usager, dans la barre d'outils, le nombre de documents qu'il lui a recommandé (processus *Navi*). A tout moment, il peut les consulter et les visiter s'il le souhaite (Figure 18). Dans cette figure, on peut voir pour chaque recommandation le nom des personnes possédant ce document et ce pour privilégier la « traçabilité des recommandations ». Une alternative, qui était proposée dans Easy-DoR était d'afficher, pour chaque recommandation, les chemins des EPA dans lesquels le document est situé (Figure 19). Un clic sur une partie du chemin d'un document permet d'accéder directement au contenu du répertoire correspondant favorisant l'exploration des EPA.

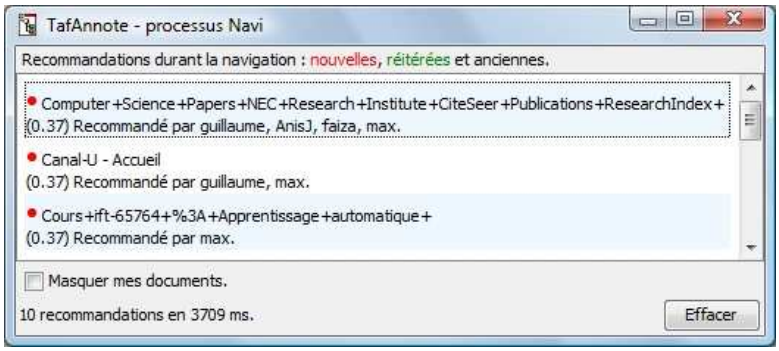


Figure 18 - Fenêtre de recommandation durant la navigation (processus *Navi*)

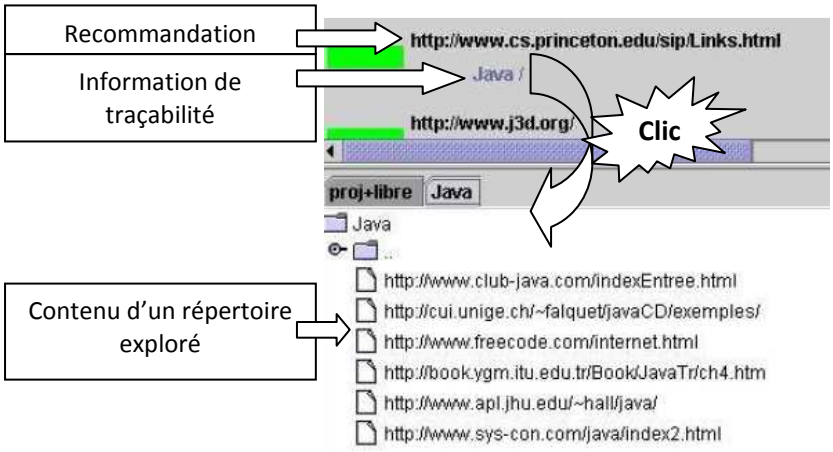


Figure 19 - Fenêtre de recommandation durant la navigation alternative (focalisation sur la traçabilité)

Lors de la navigation le système intègre directement les annotations au contenu du document auquel elles sont associées. Nous pouvons souligner l'attention particulière portée à l'expressivité de l'ancre (en comparaison avec la Figure 11). Les pictogrammes utilisés correspondent à ceux utilisés dans la fenêtre de création (Figure 16 - partie de droite), le numéro indiqué précise le nombre de réponses à l'annotation. Le filtrage des annotations est alors possible (processus Adapt-Affichage). L'exemple suivant (Figure 21) présente le résultat de l'adaptation de l'affichage en fonction de la validation sociale (et plus particulièrement des annotations qui font consensus). Dans ce cas, les annotations étant validées ✔ sont mises en valeur (taille de l'icône corrélée à la valeur de la validation sociale) par rapport à celles qui ont été réfutées socialement ✗.

Sur la base de l'EPI construit par l'usager, le processus *réorg* (disponible dans Easy-DoR) peut alors être exécuté (Figure 20). La fenêtre de gauche présente l'EPI existant de l'usager tandis que la fenêtre de droite présente la classification proposée par le système.

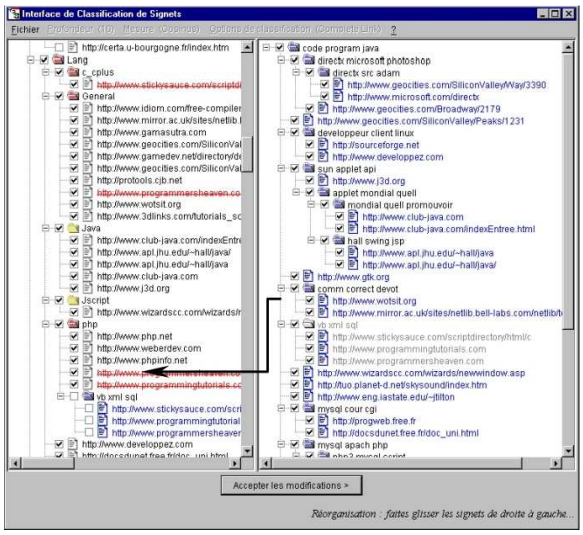


Figure 20 - Processus de réorganisation des EPI

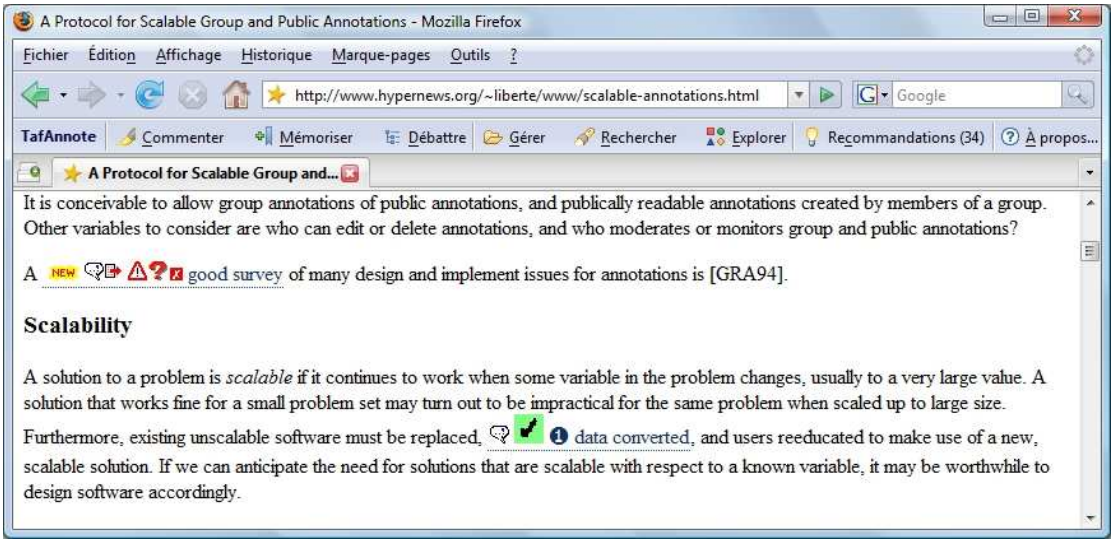


Figure 21 - Affichage des annotations et adaptation de l'affichage

Enfin, TafAnnote intègre une fonctionnalité de recherche classique par requêtage permettant aux usagers de rechercher des documents par mots-clés dans les différents EPA (Figure 22).

Figure 22 - Fenêtre de recherche dans les EPA du système

C.7.2 Point d'ancrage

Le point d'ancrage est l'un des éléments les plus importants à considérer lors de l'implantation d'un système d'annotations tel que TafAnnote. En effet, ce point d'ancrage permet au système de mémoriser le passage du document auquel l'annotation est rattachée pour pouvoir les réassocier ultérieurement (lors de la visualisation du document par exemple). Il est important que ce point d'ancrage soit robuste. Il existe différents types de points d'ancrage allant de la mémorisation du texte annoté (Denoue, 2000) à l'usage de techniques basées sur les chemins tels que *XPointer* (DeRose, et al., 2002). Cette variété de possibilités est soulignée dans l'étude comparative des systèmes d'annotations (Cabanac, et al., 2005).

Quelle que soit l'approche utilisée, les modifications apportées au contenu peuvent avoir un impact sur l'annotation (Kahan, et al., 2002) :

- Les annotations deviennent **orphelines** : le point d'ancrage ne peut plus être résolu,
- Les annotations deviennent **trompeuses** : le point d'ancrage peut être résolu mais le texte annoté n'existe plus ou a été modifié.

Dans notre cas nous avons opté pour l'usage d'un point d'ancrage robuste basé sur *XPointer* auquel nous associons la valeur de hachage du passage annoté (Cadiz, et al., 2000). Cette valeur de hachage permet au système d'identifier les annotations trompeuses.

C.7.3 Respect de la vie privée

Cet aspect est important dès l'instant où des informations relatives à l'usager sont stockées et surtout utilisées par le système. Ainsi, pour garantir le respect de la vie privée, les implantations réalisées garantissent à l'usager le contrôle de ses informations. Ainsi, à tout moment l'usager peut indiquer les informations qu'il souhaite partager ou non (niveau de visibilité des annotations - Figure 16, désactivation des processus *navi* ainsi que *reco*). Seules les informations accessibles seront donc traitées par les processus.

C.8 Bilan et résultats

L'originalité des travaux que nous avons menés réalisés au niveau macroscopique de la gestion documentaire est triple :

- **Préserver** et **amplifier** la relation entre les usagers et l'information qu'ils manipulent. Ainsi, l'approche proposée s'intègre naturellement dans l'activité réelle de l'usager et est centrée sur celle-ci,
- **Rendre accessible** les mines d'or d'informations. Par le biais des recommandations et des informations de traçabilité associées, les usagers peuvent explorer les EPA selon des points d'entrée liés à leur activité courante,
- **Améliorer l'interaction et les échanges** entre les différentes activités documentaires qui aujourd'hui restent cloisonnées.

C.8.1 Préserver et amplifier la relation entre les usagers et informations : approche centrée « usager »

La préservation et l'amplification de la relation entre les usagers et les informations sont développées dans nos propositions au travers de :

- l'usage des annotations qui sont d'ores et déjà communément exploitées. Ce concept d'annotation a été étendu en **annotation collective**,
- l'attention portée à **l'interaction** et à **l'interprétation** des informations affichées aux usagers (traçabilité dans les recommandations lors de la navigation, expressivité des ancres…),
- le caractère **adaptatif** des différents processus proposés (centrés usagers),
- la définition d'un **point d'ancrage robuste** permettant le repositionnement sans ambigüité des annotations dans un document,
- le filtrage possible des annotations pour **favoriser la lecture active** lorsque le nombre d'annotations est important dans un document,

- le **respect de la vie privée**,
- **l'implantation** sous la forme de barre d'outils s'intégrant naturellement dans un navigateur Web très utilisé par les usagers.

C.8.2 Amélioration de l'interaction entre les activités documentaires : approche centrée « activité »

L'amélioration de l'interaction entre les activités est développée dans nos travaux au travers de :

- la proposition d'une architecture intégrée qui repose sur les annotations collectives dans laquelle **cinq briques fonctionnelles** ont été structurées. Le **Erreur ! Source du renvoi introuvable.** présente une synthèse des cinq processus que nous avons proposés. Cette architecture peut être étendue assez facilement du fait de la richesse de contenu des annotations collectives pour ajouter de nouvelles fonctionnalités. L'évolution potentielle de cette architecture est donc importante, notamment dans le cadre de la RI (Cabanac, et al., 2007c).

Nom du processus	Activité(s) source(s)	Activité(s) amélioré(es)	Objectif (s)
Navi	⑥ (classement)	① (navigation), ④ (diffusion)	Recommandation de documents durant la navigation. Repose sur l'indicateur « **similarité d'organisation** ».
ProtoDoc	⑤ (Exploitation) ⑥ (classement)	② (Création)	Construction d'un document de synthèse contenant les informations liées à un ensemble d'annotations.
Reco	① (navigation)	④ (diffusion), ⑥ (classement)	Recommandation de documents directement dans l'EPA des autres individus. Les documents proviennent des navigations de tous les utilisateurs du groupe.
Réorg	⑥ (classement)	⑥ (classement)	Aide à la réorganisation des documents contenus dans l'EPA.
AdaptAffichage	-	⑤ (Exploitation)	Filtrage et Mise en valeur des annotations pertinentes pour l'usager. Repose sur les caractéristiques des annotations ainsi que sur l'indicateur « **validation sociale** **».**

Tableau 4 - Synthèse des processus proposés pour la fédération des activités documentaires

- une attention particulière apportée à la **validation** (expérimentation) des travaux permettant de vérifier leur intérêt. Certaines des validations réalisées ont nécessité la proposition et l'application d'un cadre expérimental complet (pour la validation sociale par exemple).

- la proposition de deux indicateurs centrés activités :
    - la **similarité d'organisation** qui est différente de la mesure de similarité sur le contenu,
    - la **mesure de validation sociale** qui permet d'approximer la perception humaine du consensus d'une annotation issu du fil de discussion associé.

Un point fort des propositions réside également dans le fait qu'elles dépassent les frontières des activités documentaires et peuvent être adaptées à d'autres contextes applicatifs comme le décisionnel. Nous avons étudié l'intégration des annotations collectives dans le contexte des entrepôts de données (Teste, 2000) pour favoriser les échanges entre analystes et dans le même temps la capitalisation de leur expertise ainsi que l'interprétation des données visualisées (traçabilité). Ces travaux (Cabanac, et al., 2006b), (Cabanac, et al., 2006c), (Cabanac, et al., 2007d), (Cabanac, et al., 2009b) ont notamment permis de définir un point d'ancrage robuste dans ce contexte applicable aussi bien aux modèles de l'entrepôt qu'aux tables multidimensionnelles.

En conclusion de ce chapitre, nous pouvons indiquer que les travaux menés respectent les objectifs fixés. Dans le même temps, deux leçons principales peuvent être soulignées :

- Le partage d'information (tout en respectant le respect de la vie privée) peut être considéré comme une base importante pour améliorer les activités documentaires de façon globale surtout dans un cadre organisationnel : **gestion collective d'information**,
- Les systèmes doivent donner aux usagers toutes les informations nécessaires pour pouvoir apprécier la provenance des informations qui leur sont soumises : **traçabilité**.

La validation expérimentale des différentes propositions a permis de souligner la qualité des différents processus. Cependant certaines perspectives ont été identifiées. Elles font l'objet de la section suivante.

C.9 Valorisation et support à la recherche[13]

Ces travaux servent de support à deux thèses dont une a été soutenue et un stage de Master 2. Ils ont donné lieu à 15 publications.

Encadrement	Publications				Projets associés
	RI	RN	CI/WI	CN	
2 thèses (G. Cabanac, D. Dudognon), 1 M2 (G. Cabanac)	1	2	7	5	*Overblog*

[13] Plus de détails concernant les thèses et Master 2 encadrés sont fournis en section F.1 (p. 103).

Les propositions concernant la fédération des activités documentaires autour du concept d'annotation collective ont été réalisées dans le cadre du Master 2 et de la thèse de Guillaume Cabanac.

La thèse, débutée en février 2010, de Damien Dudognon, a pour objectif, entre autres, la définition d'indicateurs basés sur l'activité humaine autour des informations dans le contexte des blogs. Le cadre d'expérimentation de cette thèse est la collaboration avec Overblog.

C.10 Perspectives relatives aux propositions

Nous pouvons identifier différentes perspectives à nos travaux relatifs à la fédération des activités documentaires et ce à deux niveaux :

- au niveau des processus proposés,
- au niveau de l'approche globale de fédération des activités documentaires.

C.10.1 Perspectives relatives aux processus proposés

L'ensemble des différents processus proposés ont été implantés, excepté le processus *ProtoDoc* qui n'est pas central à la fédération des activités documentaires. Il n'a pas encore été implanté et reste à l'ébauche de proposition. Une première perspective serait donc d'implanter et évaluer ce processus. Pour ce faire les techniques de résumé automatique de texte (Usunier, et al., 2006) et de classification pourraient être étudiées et utilisées.

Le processus *Reco* repose sur une approche uniquement basée sur le contenu des documents sauvegardés dans un répertoire de l'EPA. Ceci implique que cette approche est sensible à l'hétérogénéité du contenu des documents dans les répertoires. En effet, le classifieur actuel reposant uniquement sur la similarité de contenu ne permet pas de caractériser efficacement un ensemble de documents ayant des contenus variés (répertoire « fourre-tout » par exemple). Pour limiter cette sensibilité et permettre au processus d'optimiser les recommandations, il est nécessaire de disposer de l'information sur la raison pour laquelle un document est inséré dans un répertoire particulier. Une solution pourrait être une caractérisation des différents répertoires des EPA pour estimer une homogénéité selon différents critères (date, auteur, contenu…) et plus seulement sur la base du contenu seul. Le processus *Réorg* souffre du même problème et offre donc les mêmes perspectives.

Le processus *Navi* permet d'obtenir des documents au cours de la navigation. Cependant, les recommandations faites reposent sur la similarité d'organisation uniquement. Ainsi, si pour une navigation, aucun des documents visités ne se trouvent dans les EPA, aucun document ne sera recommandé. Une évolution de ce processus pourrait être vue en mixant la similarité de contenu et d'organisation :

- au niveau de l'identification d'un document dans le multi-arbre : si le document visité n'est pas présent, le système pourrait utiliser le document le plus similaire au document visité pour identifier les recommandations,
- au niveau de la recommandation : le système pourrait améliorer la liste de documents recommandés en y ajoutant des documents similaires en termes de contenu. Cela peut s'apparenter à un problème de fusion de résultats de recherche (Kompaore, et al., 2008).

Les différents processus reposent sur les EPA partagés des usagers du système. Pour assurer la confidentialité des informations soit les processus peuvent être désactivés (comme *Reco*), soit les droits d'accès aux informations partagées (annotations) peuvent être

mis en place. Pour le moment, ces derniers sont limités à « public » et « privé ». Une nuance plus importante dans ces **droits d'accès** serait appréciable pour favoriser l'acceptation de l'outil par les usagers. Par ailleurs, les problématiques de sécurité du système global (piratage de la base d'annotations…) n'ont pas été traitées dans nos travaux et devraient faire l'objet d'un travail à part entière.

C.10.2 Perspectives relatives à l'approche globale de fédération des activités documentaires

Les perspectives liées au processus *Reco, Réorg* et Navi, posent la question liée à **l'intention de l'usager**. Ainsi une piste d'amélioration des propositions passe par l'identification de cette intention et sur la manière de la traduire dans les processus composant l'architecture.

Par ailleurs même si nous avons considéré ce problème dans nos propositions (dans le processus **Navi** par exemple), la **traçabilité** des informations, traitées par les différents processus et présentées à l'usager, mériterait d'être améliorée. En effet, cette notion se retrouve dans les outils de GED au travers de la fonctionnalité de « *Records Management* ». Ainsi, l'usage de bases de connaissances externes et de processus complémentaires permettrait de qualifier de façon plus précise les informations présentées à l'usager. Dans ce contexte, une thèse pour laquelle je participe à l'encadrement, a débuté en février 2010. Il s'agit dans le contexte du Web et des « user-generated content » de qualifier les informations, les usagers (expertise…), les relations entre usagers et informations ainsi que les mouvements associés (création, buzz…). Cette thèse (contrat CIFRE) profite d'un cadre d'expérimentation : une plateforme en ligne de plus de 1 300 000 blogs. Au travers de cette plateforme, une mine d'informations pertinentes pour nos travaux. Ainsi par exemple une grande quantité de logs (accès, requêtes…) sont disponibles. Ces informations nous permettent d'expérimenter nos propositions dans un contexte applicatif concret.

Les deux indicateurs proposés (indicateur de validation sociale et mesure de similarité basée sur l'organisation) ont été intégrés dans des processus (*AdaptAffich*, *Navi*) pour l'amélioration de la gestion documentaire. Une réflexion autour de nouveaux indicateurs liés à l'activité des usagers (pertinence sociale d'une information, buzz lié à une information…) devrait également permettre d'améliorer l'architecture proposée.

C.11 Perspectives relatives aux activités documentaires

Aujourd'hui la réalisation des activités documentaires est essentiellement individuelle. Nos propositions visent à la rendre collective dans le sens où les systèmes supportant les activités documentaires partagent des informations essentiellement de façon implicite. Cependant, afin d'optimiser encore plus la dynamique autour des usagers et des activités documentaires, une perspective est de développer leur **dimension collaborative**. Ainsi certaines activités et notamment la RI pourraient intégrer une collaboration synchrone entre les usagers. Une première série de travaux dans le cadre la RI est montré dans (Foley, et al., 2008).

Après la présentation de nos propositions centrées sur les activités documentaires (niveau macroscopique), nous présentons les travaux que nous avons réalisés au niveau microscopique c'est-à-dire dans les travaux visant à améliorer l'activité de recherche d'information par requêtage (systèmes de recherche d'information) à proprement parler. Ces travaux sont toujours centrés sur l'usager.

D. AMÉLIORATION DE LA RI : SRI CENTRÉS USAGERS

D.1 Introduction

Les systèmes de recherche d'information (appelés aussi moteurs de recherche dans le contexte du Web), notés par la suite SRI, sont sans doute l'un des moyens les plus utilisés par les usagers pour trouver de l'information. Cependant, du point de vue de l'usager, utiliser ces outils peut ne pas se révéler aussi simple qu'il n'y paraît comme nous l'avons souligné précédemment. Ce chapitre considère les SRI comme étant une boîte noire et se focalise principalement sur les éléments d'interaction entre l'usager et le système afin d'identifier les difficultés que rencontre l'usager. Ainsi si l'on reprend les quatre étapes de la recherche d'information (section B.1.1), les travaux présentés dans ce chapitre sont relatifs à :

- La formulation des besoins,
- L'accès aux documents retrouvés et en particulier la présentation des résultats de recherche à l'usager.

Nous nous sommes donc intéressés à ces deux dimensions de la recherche d'information dans nos travaux avec pour objectif d'apporter des réponses aux problèmes auxquels l'usager est confronté.

D.1.1 Formulation des besoins

La formulation de requête est une étape cruciale dans le processus de RI. En effet, d'après une étude de M. Sanderson (Sanderson, 2008), entre 7% et 23% des requêtes sur le Web sont ambigües. Afin d'aider l'usager à mieux formuler ses besoins, différentes stratégies peuvent être mises en place :

- la reformulation de requêtes,
- la construction visuelle et incrémentale de requête.

D.1.1.1.1 La reformulation de requêtes

L'objectif de la reformulation de requêtes est d'améliorer la qualité de la requête en modifiant les mots qui les composent. Ainsi, la reformulation de requêtes repose généralement sur :

- La technique classique de reformulation par **réinjection de pertinence** (B.2.2.3.1),
- La **désambigüisation de requête** qui consiste à utiliser une base de connaissances externe (relative à des liens notamment sémantiques entre les termes par exemple) pour contextualiser la requête en ajoutant des termes par exemple (Jing, et al., 1994). Cette adaptation de la requête repose essentiellement sur des informations externes (thésaurus, ontologies (Hoang, et al., 2008)…) ou encore sur l'EPI de l'usager (Bottraud, et al., 2004) ou encore sur des tags (Broder, et al., 2008),

- Une **approche mixte** telle que celle proposée dans *WebCluster* (Harper, et al., 1999). Ce système basé sur une *approche par médiation* (échanges entre le système et l'usager) laissant l'usager maître de la situation permet à l'usager d'améliorer sa requête. Le système, à partir d'une requête originale, propose à l'usager de choisir parmi un ensemble de groupes de documents ceux qui se rapprochent de ses besoins. Ces groupes de documents sont construits sur la base d'une base de connaissances externe. A partir de la sélection de l'usager, le système génère une requête qu'il renvoie à l'usager ou qu'il soumet à un moteur de recherche.

L'orientation actuelle des travaux dans le domaine se focalisent de plus en plus sur la désambigüisation de requêtes du fait que de plus en plus de bases de connaissances externes aux documents sont exploitables, notamment dans le cadre du Web 2.0, telles que Wikipédia…

D.1.1.1.2 Construction visuelle et incrémentale de requête

Comme alternative à la désambigüisation de requête, certains systèmes proposent à l'usager une aide à la construction de requête. Cette aide repose essentiellement sur la présentation des relations entre les termes. Ces relations peuvent être sémantiques ou encore correspondre à la corrélation qu'entretiennent des termes dans le log de requêtes d'un moteur de recherche (Chevalier, et al., 2009c), GoogleWonderWheel[14], Entity Association Graph[15]. Cette aide permet ainsi à l'usager de construire de façon incrémentale sa requête en parcourant les termes utilisés par les autres usagers du moteur. Au regard du nombre de termes possibles, ces aides sont généralement fournies sous la forme d'interfaces graphiques.

D.1.1.1.3 Recherche à facettes (« faceted search »)

Pour aider l'usager à formuler ses besoins et ainsi se focaliser le plus rapidement possible sur les informations qui répondent le mieux à ses besoins, certains systèmes proposent une recherche à facettes (« faceted search ») en complément de l'usage d'une requête. La recherche à facettes permet à l'usager d'accéder aux documents qui l'intéressent à partir d'un ensemble de critères (facettes). L'idée sous-jacente est qu'un objet complexe peut être observé selon différentes perspectives ou facettes (Kwasnik, 1999). Les différentes facettes utilisées permettent d'obtenir l'équivalent d'une classification multiple des documents. L'avantage principal de cette approche est que la diversité des facettes permet à l'usager de s'orienter rapidement sur les informations qu'il désire contrairement à une classification unique (Zamir, et al., 1999). L'efficacité des facettes pour la RI a été démontrée dans (Kules, et al., 2009).

A l'autre extrémité du processus de RI, la présentation des résultats de recherche est le deuxième niveau dans lequel l'usager est fortement impliqué.

[14] http://www.googlewonderwheel.com/

[15] http://adlab.microsoft.com/Vnext/Entity-Association-Graph/

D.1.2 Visualisation pour la RI

En réponse à une interrogation de la part d'un usager, un SRI retourne un ensemble de résultats qui peut contenir plusieurs milliers (voire millions) de documents selon les contextes. La difficulté pour l'usager est alors d'identifier dans ces résultats ceux qui répondent à ses besoins. La visualisation des résultats de recherche a donc pour objectif d'aider l'usager dans l'interprétation des résultats et dans la localisation des documents pertinents.

Dans ce cadre, la visualisation des résultats sous la forme d'une liste ordonnée selon le critère de pertinence système est la technique utilisée par la majorité des SRI (Figure 23).

Figure 23 - Exemple de visualisation sous la forme d'une liste de résultats (http://www.google.fr)

Cependant, bien que ce type de visualisation ait l'avantage d'être simple à générer par le système et simple pour la plupart des utilisateurs de par sa facilité d'interprétation, il possède quelques inconvénients majeurs tels que :

- l'absence de relation réelle entre le contenu des différents documents retrouvés. En effet, aucune relation sémantique (concernant le contenu des documents retrouvés) entre les documents n'est proposée. Ainsi, l'utilisateur doit réaliser tous les efforts afin de comprendre le contenu des documents retrouvés en les visitant un à un ;
- la désorganisation thématique des résultats. En effet, les résultats appartenant à différentes thématiques sont mélangés dans la liste des résultats en fonction de leur pertinence système ;
- l'absence de vue d'ensemble des résultats. Seulement une dizaine de résultats sont généralement affichés simultanément à l'écran. Il n'est donc pas possible, pour l'utilisateur, d'avoir un aperçu de l'ensemble des réponses.

En réponse à ces limites, un grand nombre de travaux dans le domaine de la visualisation des résultats de recherche ont été proposés dans la littérature mais également mis en œuvre dans différents systèmes disponibles sur Internet (Chen, 2006), (Bonnel, et al., 2008), Visual Search Engine[16].

Pour essayer d'organiser ces travaux, nous avons identifié deux grands courants de taxonomies des interfaces pour la RI :

- les taxonomies qui reposent sur le type de métaphore utilisée (*e.g.* visualisation en arbres ou en listes) (Chi, 2000), (Hascoët, et al., 2001),

[16] http://www.masternewmedia.org/top-visual-search-engines-the-most-interesting-ways-to-visually-explore-search-engine-results/

- les taxonomies qui reposent sur des caractéristiques des informations (par exemple le nombre de documents, les liens entre les documents ou la pertinence) (Zamir, 1998), (Bonnel, 2006).

Nous sommes personnellement sensibles à la deuxième taxonomie qui privilégie les informations visualisées par rapport à la façon dont elles sont visualisées. En effet, cette taxonomie nous semble plus exploitable du point de vue d'un concepteur de SRI puisqu'elle permet d'accéder aux interfaces à partir des informations que l'on souhaite visualiser. La Figure 24 présente la taxonomie sur laquelle repose nos travaux. Les noms des interfaces de visualisation que nous avons considérées comme représentatives sont indiquées sous le nom des catégories. Une liste des travaux existants et une présentation détaillée de la taxonomie sont disponibles dans (Bonnel, et al., 2008).

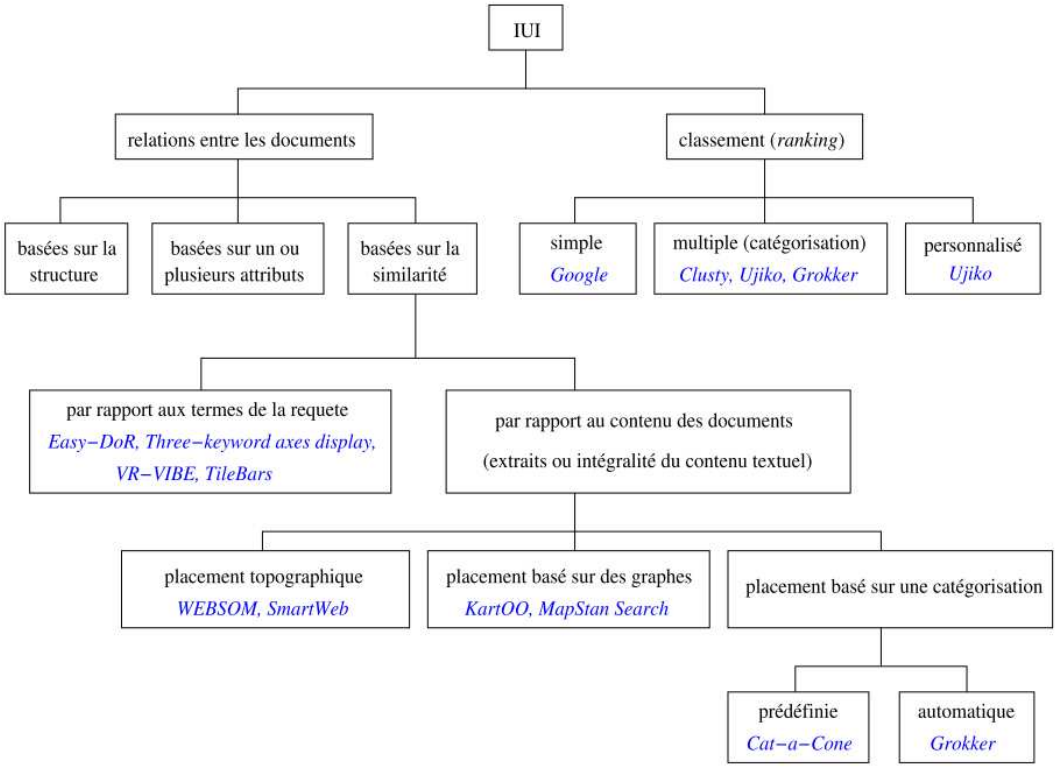


Figure 24 - Taxonomie des visualisations pour la recherche d'information (Bonnel, et al., 2008)

Cette taxonomie souligne la grande variété dans les possibilités de visualisation des résultats de recherche. Cependant, avant de pouvoir utiliser une visualisation, il est nécessaire de l'évaluer. L'évaluation d'une interface a le plus souvent pour objectif de vérifier si les aspects cognitifs (l'utilisabilité notamment) sont respectés permettant ainsi à l'usager de l'utiliser pleinement. Deux grandes tendances d'approches d'évaluation ont été développées :

- les méthodes **analytiques** : elles sont basées sur la simulation des activités de l'utilisateur, sans son implication. Cette simulation est effectuée par des experts ;
- les méthodes **empiriques** : elles ont basées sur l'observation des comportements et des attitudes des utilisateurs en situation d'utilisation de l'interface à évaluer.

D.2 Enjeux & Objectifs de nos travaux

L'objectif des travaux relatifs à l'amélioration centrée usagers des SRI que nous avons menés est triple :

- Faire que la **formulation des besoins** soit facilitée pour l'usager. L'objectif est d'apporter une aide à la formulation des besoins sur la base d'une approche par médiation (allers-retours entre l'usager et le système),
- Favoriser l'exploitation des résultats de recherche par l'usager en proposant **une ou plusieurs visualisations adaptées** à ses besoins,
- Permettre à l'usager de **reprendre le contrôle du processus de RI** et que l'activité (notamment cognitive) demandée à l'usager soit limitée et répercutée sur le SRI. Par ce biais, les usagers gagneront en efficacité et en confort alors qu'aujourd'hui c'est l'usager qui fait au final le plus gros du traitement.

Dans nos propositions présentées dans les sections suivantes, les techniques de visualisation revêtent un intérêt certain au même titre que la dimension collective.

D.3 Aide à la formulation des besoins : Désambigüisation de requêtes par médiation

Ce premier axe de travaux vise à apporter une aide à la formulation des besoins qui n'est pas une étape évidente pour les usagers. Notre proposition repose sur une exploitation collective des informations liées aux requêtes : les logs de requêtes. L'intérêt d'exploiter ces informations est qu'elles permettent, fidèlement à notre approche centrée usagers et activités, de partager l'expérience et l'activité de recherche entre les différents usagers du système de recherche. Ces informations sont peu disponibles (Shingo, et al., 2006) car de ce fait elles sont une mine d'or du point de vue commercial pour les systèmes de recherche (revente de mots-clés par exemple). Dans nos travaux, nous avons exploité l'historique de requêtes fourni (par erreur) par le moteur aol.com (20 millions de requêtes de 650000 usagers sur 3 mois) et disponible désormais à des fins de recherche à l'URL http://gregsadetsky.com/aol-data.

Sur cette base, nous proposons une approche de désambigüisation de requête par médiation. L'objectif de cette approche est de permettre à l'usager d'obtenir une requête correspondant à ses besoins et de comprendre comment modifier sa requête.

L'idée sous-jacente de ces travaux vient directement du constat que le nombre de mots-clés soumis aux SRI est très faible (Silverstein, et al., 1998), (Jansen, et al., 2000), (Spink, et al., 2002). De ce fait la qualité des résultats de recherche s'en ressent (Voorhees, et al., 1999). Pour permettre à l'usager de rester maître de sa requête tout en lui apportant une aide précieuse, nous avons choisi une approche basée sur un principe proche de la médiation.

Ainsi à partir d'une requête initiale, l'objectif de nos travaux est d'identifier des groupes cohérents de termes (agrégats) parmi lesquels l'usager pourra sélectionner les termes qui lui conviennent et ainsi modifier sa propre requête avant de la soumettre réellement au SRI. Cette approche est assez proche des travaux menés dans (Mechkour, et al., 1998) ou encore (Airio, et al., 2004) à la différence que nous proposons à l'usager des agrégats de termes plutôt que des classes de documents ou des ontologies plus difficiles à maintenir. Chaque agrégat de termes peut être assimilé par certains aspects à un contexte d'usage des termes de la requête. Ainsi, sur la base de ces agrégats et de l'importance relative des termes dans ces agrégats, l'usager est capable d'identifier les contextes qui l'intéressent et peut ainsi par

lui-même désambigüiser sa requête. Ces agrégats de termes sont construits à partir des mots utilisés dans les requêtes présentes dans l'historique de requêtes du moteur de recherche (de aol.com dans notre cas).

De nos motivations, nous avons pu extraire deux contraintes principales qui se sont imposées naturellement à notre travail :

- (C1) les agrégats construits doivent être cohérents (c'est-à-dire que les termes qui composent l'agrégat doivent être liés),
- (C2) un terme doit pouvoir appartenir à plusieurs agrégats puisqu'il peut avoir plusieurs sens et donc être utilisé dans différents contextes.

Les approches présentées précédemment notamment de clustering reposent essentiellement sur des approches de partitionnement ne permettant pas de respecter (C2). Nous avons donc élargi l'étude des travaux basés sur l'identification de communautés dans les graphes (Fortunato, 2010). Cependant, la difficulté est que dans notre cas le graphe de termes que l'on peut construire sur la base des requêtes du fichier historique d'AOL est un graphe dans lequel les arêtes sont orientées et pondérées. En effet, l'importante d'une relation entre deux termes n'est pas symétrique et chaque couple de termes n'a pas la même probabilité d'apparition dans les requêtes. De ce fait, la plupart des démarches applicables aux graphes sont difficiles à mettre en place ou ne sont pas assez souples pour nous permettre d'adapter convenablement leur fonctionnement pour atteindre notre objectif. Nous nous sommes finalement orientés vers des travaux, toujours sur les graphes, qui reposent sur des approches de résolution de contraintes géométriques. Un système de contraintes géométriques se compose d'un ensemble d'objets géométriques soumis à des contraintes géométriques. Résoudre un système de contraintes géométriques consiste à fournir une position, une orientation et des dimensions à chacun de ses objets géométriques de sorte que toutes les contraintes géométriques soient satisfaites. Nous nous sommes ainsi orientés vers les travaux d'Hoffman *et al.* (Hoffman, et al., 1997) et de Jermann *et al.* (Jermann, 2002) qui présentent la particularité d'offrir la possibilité d'utiliser notamment un opérateur d'extension pour les adapter au contexte applicatif.

La proposition que nous avons formulée (Belbèze, et al., 2009), (Belbèze, et al., 2009c) consiste donc à adapter les travaux précédents à notre contexte en y insérant les contraintes spécifiques à notre cadre applicatif. Ainsi dans notre contexte, l'opérateur a été redéfini pour prendre notamment en considération l'importance relative de l'utilisation d'un terme avec un autre. Pour apporter une robustesse et une rigidité naturelle aux agrégats construits nous nous sommes basés sur la notion de « cliques » (Luce, et al., 1949).

L'algorithme de construction proposé permet ainsi de générer des agrégats qui sont intuitivement cohérents. La Figure 25 présente les agrégats obtenus et contenant le mot « apple ». On peut ainsi voir quatre contextes d'usage de ce mot : (1) le fruit, (2), la fleur, (3) la marque informatique, (4) le cidre. Notons que seuls les mots les plus importants des agrégats sont représentés sur cette figure par soucis de lisibilité.

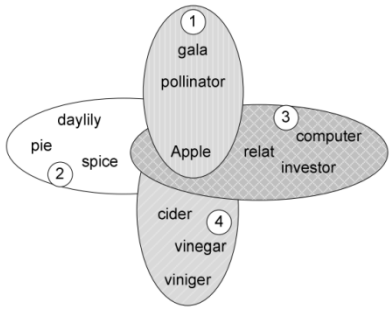


Figure 25 - Exemple de quatre agrégats obtenus pour le terme "Apple"

L'originalité de l'approche proposée est qu'elle :

- Repose sur une approche de **médiation**, c'est-à-dire que le système ne remplace aucunement l'usager. Il lui donne les clés (informations relatives aux contextes d'usages des termes de sa requête) pour construire lui-même sa requête,
- Suit **l'évolution des terminologies** du fait qu'elle repose sur les requêtes réellement soumises à un SRI. Ainsi, les contextes d'utilisation de termes évoluent en temps réel (au temps de traitement près). Cette approche repose donc sur des **communautés dynamiques de termes**.

Pour vérifier la cohérence de façon expérimentale, une première méthode empirique a été proposée (Belbèze, et al., 2009b). Cette évaluation repose sur l'hypothèse que si une requête est cohérente, elle doit permettre au SRI de retourner des documents. Une illustration de cette hypothèse est fournie dans le Tableau 5.

ID	Requête	nb de sites retournés
1	+besancenot +état +france	164 000
2	+épinard +crème +beurre	37 400
3	+saxophone +selmer +jazz	66 300
4	+selmer + besancenot +épinard	0

Tableau 5 - Exemples de la relation entre cohérence de la requête et nombre de documents retournés par le SRI

Les premiers résultats obtenus confirmaient le fait que les agrégats étaient cohérents sur différentes journées. Cependant, en poursuivant les expérimentations, nous avons pu conclure que l'approche même de validation empirique que nous avons proposée ne permettait pas de conclure totalement à la cohérence des agrégats. En effet, les résultats obtenus sur les différentes journées du fichier historique utilisées dans l'évaluation ne sont finalement pas transposables sur d'autres journées. Une nouvelle expérimentation, sur un corpus disponible dans TREC, est en cours de définition.

D.4 SRI et Visualisation d'Information

La visualisation en RI est un élément important car, comme nous l'avons souligné au travers de la revue de la littérature, elle permet d'améliorer l'accès aux documents retournés par le SRI.

Dans ce contexte, une première proposition que nous avons faite consiste à visualiser les documents retrouvés selon une a**pproche à facettes**. Elle correspond en quelque sorte à la convergence des techniques de visualisation et des approches de « faceted search ». Elle permet d'explorer différents objets, tels que des documents, selon différents points de vue.

Dans un second temps, au regard de la littérature et de la difficulté de choisir la « bonne interface » permettant une exploitation optimale du résultat de recherche, une seconde proposition consiste à mesurer l'adéquation entre un contexte de recherche et une interface de visualisation de résultats de recherche. A partir de cette adéquation, il sera alors possible de choisir les interfaces permettant une exploitation optimale des résultats de recherche.

D.4.1 Interface de visualisation multi-facettes

Suite aux travaux que nous avons menés dans le cadre de la proposition d'une interface de visualisation des résultats de recherche d'information (Chevalier, 2002), l'approche de visualisation multi-facettes que nous proposons (Cabanac, et al., 2009c) repose sur les notions de :

- **Facettes** : caractéristiques d'un objet (exemple : une collection de documents) que l'on souhaite observer,
- **Vues** : représentations d'un objet sur la base de différentes facettes permettant ainsi d'avoir une vision globale de l'objet.

La notion de facette permet donc de présenter différents aspect de l'objet visualisé. Dans le contexte de la RI on peut par exemple vouloir apprécier les documents retrouvés par au moins trois facettes : les *mots* qu'ils contiennent, les relations entre les documents à partir de la *similarité basée sur le contenu* ou encore de *la similarité basée sur l'organisation* que nous avons proposée dans le chapitre précédent. La cohabitation de ces trois facettes dans la même vue permet à l'usager d'apprécier aussi bien le contenu des documents que les relations qu'ils entretiennent entre eux. Chaque facette pour une vue donnée peut être présentée sous la forme de différentes visualisations permettant à l'usager de choisir celle qu'il juge être la plus adaptée. Ainsi, pour apprécier la similarité entre des documents basée sur le contenu, l'interface pourra utiliser une visualisation soit sous la forme d'un arbre de documents résultant par exemple d'une classification hiérarchique ascendante, soit sous la forme d'une carte de Kohonen (Kohonen, et al., 2000).

L'interface proposée peut être schématisée au travers de la Figure 26. Dans cet exemple, chaque vue est caractérisée par 4 facettes.

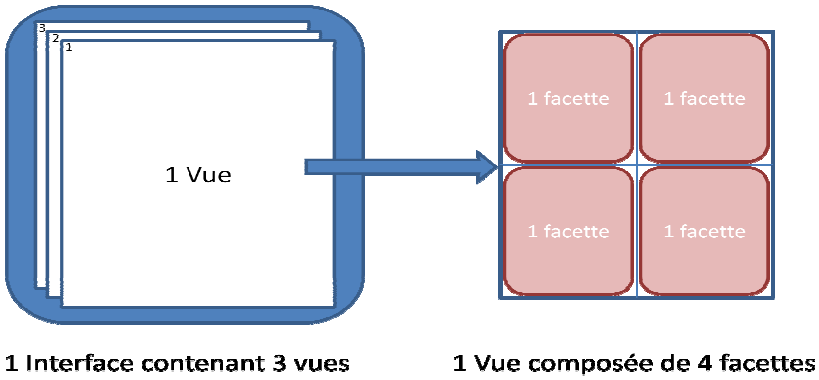


Figure 26 - Schéma général de l'interface multi-facettes

A titre d'exemple, nous présentons une liste de facettes et de vues dans le cadre de la RI. Dans cet exemple, nous avons fait le choix de considérer des facettes qui permettent à l'usager de comprendre notamment les relations qui existent entre usagers, entre usagers et informations et entre informations. Ce choix résulte d'une l'analyse de l'importance de ces relations pour la RI dans un cadre organisationnel (Canut, et al., 2005), (Canut, et al., 2005b), (Boyer, et al., 2007). Ainsi, nous avons proposé les principales facettes qui sont :

- Le Contenu (mots),
- La Similarité entre objets basée sur le contenu,
- La Similarité entre objets basée sur l'organisation,
- Les Méta-données (titre, URI, nom …).

Ces facettes peuvent être utilisées pour décrire différentes vues (couvrant les différents niveaux d'analyse) correspondant à :

- Un ensemble de documents,
- Un ensemble d'usagers,
- Un document unique,
- Un usager unique.

Le changement de vue, en plus d'être manuel, est automatique en fonction de la sélection qui est réalisée par l'usager dans les différentes facettes. Par exemple, si l'usager sélectionne dans la vue « ensemble de documents », un document unique, l'interface proposera la vue « document unique » relative à cette sélection. Dans le même temps, conformément à l'idée du « faceted search », les sélections successives agissent comme des filtres qui permettent à l'usager de se focaliser sur les informations qui l'intéressent. Ainsi si dans la vue « ensemble de documents », l'usager sélectionne un sous-groupe de documents, la vue courante sera rafraichie pour n'afficher que les informations relatives à ce sous-groupe de documents. Ces travaux se rapprochent donc des travaux de (Reiterer, et al., 2001). Tout comme ces travaux, l'interface repose sur une synchronisation des visualisations (toute sélection dans une des vues se répercute dans les autres). Cependant dans cette approche chaque visualisation est une alternative aux autres. Dans notre cas, la notion de vue composée de facettes est importante car elle permet d'obtenir une vision globale de l'objet dans la même interface que l'on ne peut pas obtenir entre différentes visualisation « indépendantes », c'est-à-dire affichées l'une après l'autre.

En conclusion, l'approche de visualisation multi-facettes proposée est originale par le fait qu'elle :

- repose sur une **complémentarité** entre les vues qui permet par exemple d'explorer les différents objets d'intérêt (documents, usagers…) et ce au travers d'une interface multi-facettes unique,
- permet une **focalisation** plus rapide de l'usager sur les objets qui l'intéressent. En effet, selon la même orientation que le « faceted search », les sélections successives de l'usager permettent de filtrer les objets disponibles dans les différentes vues synchronisées,
- favorise la **complétude** de la description de chaque vue au travers des différentes facettes proposées,
- soit **extensible** car les facettes ainsi que les vues peuvent être définies ou adaptées pour chaque contexte d'application. De plus, différentes visualisations possibles peuvent être intégrées à l'interface multi-facettes permettant de donner la possibilité à l'usager de choisir, pour une facette, la visualisation qui lui convient le mieux,

- soit « **multi-usage**s » car elle peut être utilisée pour explorer aussi bien le capital documentaire en sommeil dans les organisations (Canut, et al., 2004) que des résultats de RI.

Nous venons de présenter une orientation de nos travaux quant à une approche résultat de la convergence du principe du « faceted search » et des visualisations. Pour chaque facette, une ou plusieurs visualisations peuvent être utilisées. Cependant, à la question : quelle est la visualisation la plus adaptée pour visualiser les relations de similarité de contenu entre des documents ? Différentes réponses peuvent être données : un graphe, une carte de Kohonen, un arbre… Le deuxième pan de nos travaux concernant la visualisation dans le domaine de la recherche d'information apporte des éléments de réponses à cette question.

D.4.2 Efficacité des interfaces de visualisation

Nous avons souligné, que ce soit au travers des travaux existants ou de nos propres travaux, qu'un grand nombre de visualisations étaient proposées dans le contexte de la RI. Ces différentes visualisations ont fait l'objet d'évaluations permettant de vérifier les aspects cognitifs. Cependant, lorsqu'un concepteur souhaite mettre en place une visualisation sur un SRI, une question reste sans réponse et ce malgré les évaluations actuelles (Chevalier, et al., 2005) : *quelle(s) visualisation(s) permet(tent) d'obtenir la plus grande efficacité pour l'usager dans sa recherche ?*

Dans le cadre de nos travaux nous traduisons **efficacité** pour l'usager par :

- la pertinence et la complétude (par rapport à ses besoins) des documents qu'il a retrouvés,
- la rapidité et la satisfaction avec lesquelles il a retrouvé les documents.

Ainsi, même si quelques campagnes d'expérimentations telles que celles proposées dans (Fekete, et al., 2004), (Plaisant, et al., 2008) existent, la comparaison de l'efficacité des différentes interfaces dans le cadre de la RI reste difficile. Les campagnes d'évaluation TREC[17], INEX[18] ou encore CLEF[19] ont également proposé des évaluations incluant les interfaces de visualisation notamment au travers des tâches interactives. Cependant, ces campagnes évaluent les interfaces de façon intégrée aux SRI proposés par les participants. Ainsi, chacune des interfaces est couplée à un SRI différent. Ceci implique que la mesure de l'efficacité des interfaces est noyée dans celle du SRI lui-même.

Dans ce cadre, nous proposons une approche permettant de donner des éléments de réponse. L'objectif de nos travaux consiste à mesurer l'efficacité d'une interface de visualisation dans différents contextes de recherche. En effet, différents critères influent sur l'efficacité d'une visualisation, la variabilité humaine en est un bon exemple.

En réponse à cette problématique, nous avons développé une démarche complète et extensible de mesure de l'efficacité des interfaces de visualisation pour la RI. Cependant, la

[17] http://trec.nist.gov/

[18] http://inex.is.informatik.uni-duisburg.de/

[19] http://www.clef-campaign.org/

notion d'efficacité étant très contextuelle, nous avons tout d'abord défini le **contexte de recherche** au travers duquel elle pourra être mesurée.

D.4.2.1 Définition d'un contexte de recherche

Différents éléments peuvent être identifiés comme conditionnant la réussite d'une recherche et plus largement son efficacité, parmi lesquels :

- les connaissances des usagers (cf. B.2.1),
- le type de recherche (Rosenfeld, et al., 1998) induisant un traitement du résultat de recherche spécifique de la part de l'usager. Quatre types de recherche ont été identifiés : recherche d'une information connue, recherche d'une information quelconque, recherche exploratoire, recherche exhaustive,
- les caractéristiques de l'ensemble de résultats de recherche retourné par le système. (Reiterer, et al., 2001) souligne que par exemple le nombre et le type des documents ont un impact sur l'efficacité.

Sur la base de ces éléments, nous définissons un **contexte de recherche** autour d'un triplet <U, S, T> caractérisant l'usager (U), le système de recherche (S) et la tâche de recherche (T). Chacun de ces éléments est associé à un ensemble de caractéristiques (Bonnel, et al., 2006), (Bonnel, et al., 2006b) dont la plupart sont déjà exploitées dans le cadre de la RI :

- l'usager est décrit au travers de ses connaissances (du domaine, pratique),
- le système est décrit au travers des caractéristiques décrivant la liste de résultats de recherche (nombre de documents, hétérogénéité thématique du résultat de recherche, Précision du résultat de recherche…),
- la tâche de recherche est décrite par le type de recherche.

Pour un contexte donné, l'efficacité d'une interface de visualisation peut alors être mesurée. Pour ce faire un protocole complet a été proposé.

D.4.2.2 Protocole de mesure de l'efficacité d'une interface dans un contexte spécifique.

Le protocole repose sur un élément central qui sert de chef d'orchestre des expérimentations. Il est basé sur *un moteur de recherche virtuel* ainsi sur lequel se connectent les interfaces à expérimenter pour notamment obtenir les résultats de recherche à visualiser. Le qualificatif « virtuel » indique qu'il ne s'agit pas là d'un moteur de recherche à proprement parler mais d'un système simulant le fonctionnement d'un moteur (les requêtes ainsi que les résultats de recherche renvoyés aux interfaces sont prédéfinis). Le fait d'utiliser un système de recherche « virtuel » unique pour toutes les interfaces permet de :

- maîtriser les résultats renvoyés à chaque interface. Le moteur assurant la gestion des listes de résultats de recherche, il est possible de répliquer une expérimentation en envoyant plusieurs fois la même liste de résultats de recherche à différents usagers pour la même interface ou à différentes interfaces. Du même coup, le système peut mesurer les caractéristiques de ces listes de résultats lui permettant ainsi de valoriser la partie Système du contexte de recherche,
- comparer les différentes interfaces sur une même base permettant ainsi de se focaliser uniquement sur l'efficacité de chaque interface.

Chaque interface obtient donc, conformément à la dimension Système du contexte de recherche, une liste de résultats de recherche que chaque usager « testeur » doit explorer pour identifier les documents pertinents pour la requête qui lui a été confiée.

A la fin de la recherche, l'usager indique s'il pense avoir atteint ou non les objectifs fixés. Pour vérifier ce point, l'usager doit également fournir la liste ordonnée des documents qu'il juge comme pertinents pour que l'on puisse comparer ces documents avec ceux que le système sait être pertinents. En complément, nous conservons des caractéristiques secondaires telles que le temps d'affichage de l'interface ainsi que le temps de recherche total (entre le moment où l'interface affiche les résultats et le moment où l'usager termine sa recherche).

Grâce aux valeurs des caractéristiques liées aux différents contextes et aux résultats de recherche, un arbre de décision est construit pour chaque interface par l'algorithme C4.5 (Quinlan, 1993) par exemple. Cet arbre de décision (Figure 27), pour lequel la conclusion est la réussite ou non de la recherche, permet de comprendre dans quel cas l'interface a permis ou non d'atteindre les objectifs de recherche. Les valeurs de *support* et de *confiance* associées à chaque règle dans l'arbre de décision complètent cette interprétation.

Par exemple, d'après la Figure 27, nous pouvons souligner l'inefficacité de l'interface évaluée dans un contexte spécifique. Ainsi, en parcourant l'arbre on trouve la règle :

> ***si*** *la précision (Prec) est faible (Low) (c'est-à-dire que la proportion de documents pertinents dans l'ensemble de documents est faible)*
> ***et*** *que le nombre de documents (NbDocs) est supérieur ou égal à 30*
> ***et*** *que la connaissance du domaine de recherche (Domain) est faible (Néophyte)*
> ***alors*** *la recherche conduit à un* ***échec*** *(no)*.

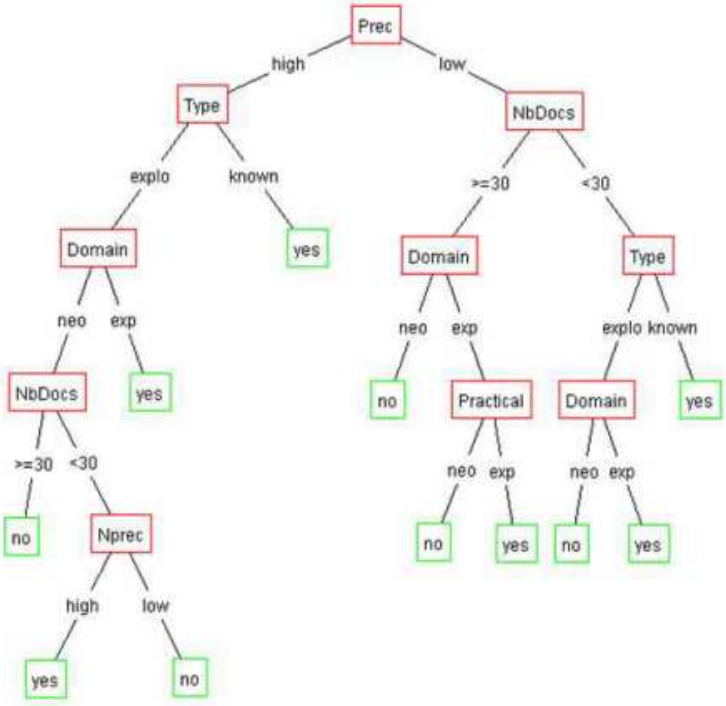


Figure 27 - Exemple d'arbre de décision soulignant les contextes dans lesquels l'interface a conduit à un succès (yes) ou à un échec (no). Les valeurs de support et confiance ne sont pas précisées.

Ainsi, grâce à ce protocole d'expérimentation, nous sommes en mesure de mettre en évidence les éléments du contexte de recherche qui conditionnent l'efficacité de chaque interface de visualisation.

En conclusion, la particularité de nos travaux (Bonnel, et al., 2008b) réside dans le fait :

- qu'elle repose sur un **protocole complet** basé sur la caractérisation de contextes de recherche dans lesquels l'efficacité de n'importe quelle interface de visualisation pour la RI est mesurée,
- que les **expérimentations soient réplicables** et réalisées sur une **base commune** (grâce au moteur de recherche virtuel qui maîtrise les documents envoyés aux interfaces) favorisant la comparaison des interfaces de visualisation,
- que les expérimentations peuvent être **réalisées en dehors du cadre d'un laboratoire et ce par n'importe qui**. En effet, une plateforme de mesure de l'efficacité des interfaces a été implantée. Celle-ci est fournie clé en main en ligne et est associée avec une API (*Application Programming Interface*) permettant à n'importe quelle interface d'évaluer sa propre efficacité. Cette plateforme est présentée dans la section suivante,
- que le protocole permet la **comparaison** des notions subjectives de succès ou d'échec pour un usager avec les mesures de qualité (statistiques) d'une liste de résultats de recherche développées en RI (précision, rappel, F-mesure…).

D.4.3 Prototypes

Les différentes propositions faites dans le cadre de la visualisation pour les SRI ont été implantées dans deux prototypes : une interface multi-facette et une plateforme de mesure de l'efficacité des interfaces de visualisation des résultats de recherche.

D.4.3.1 Interface de visualisation multi-facettes

L'interface multi-facettes que nous avons implantée repose sur les vues et facettes présentées à titre d'exemple dans la section D.4.1. Cette implantation est plutôt destinée à un SRI organisationnel car elle repose sur les EPA et particulièrement sur la mesure de similarité basée sur l'organisation (C.6.1). Cette interface ainsi que son fonctionnement sont présentés au travers des copies d'écran suivantes.

La Figure 28 présente l'interface multi-facettes et plus particulièrement la vue 1 « Ensemble de documents » dans laquelle nous pouvons identifier les différentes facettes (signalées par *fx.y* où x correspond au numéro de la vue et y le numéro de la facette dans la vue). La taille relative des facettes dans une vue peut être adaptée par l'usager.

Nous avons souligné le fait que plusieurs visualisations pouvaient être proposées dans une vue pour une même facette. La Figure 29 présente la même vue que la figure précédente à la différence que la première facette (f1.1) n'est plus un arbre mais une carte de Kohonen.

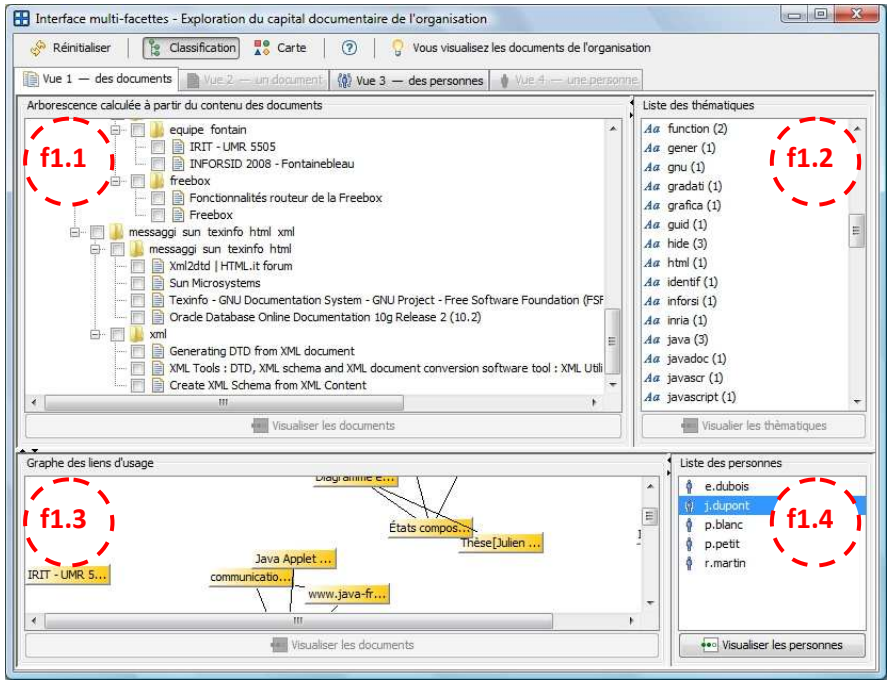


Figure 28 - Vue 1 de l'interface multi-facettes proposée pour la RI

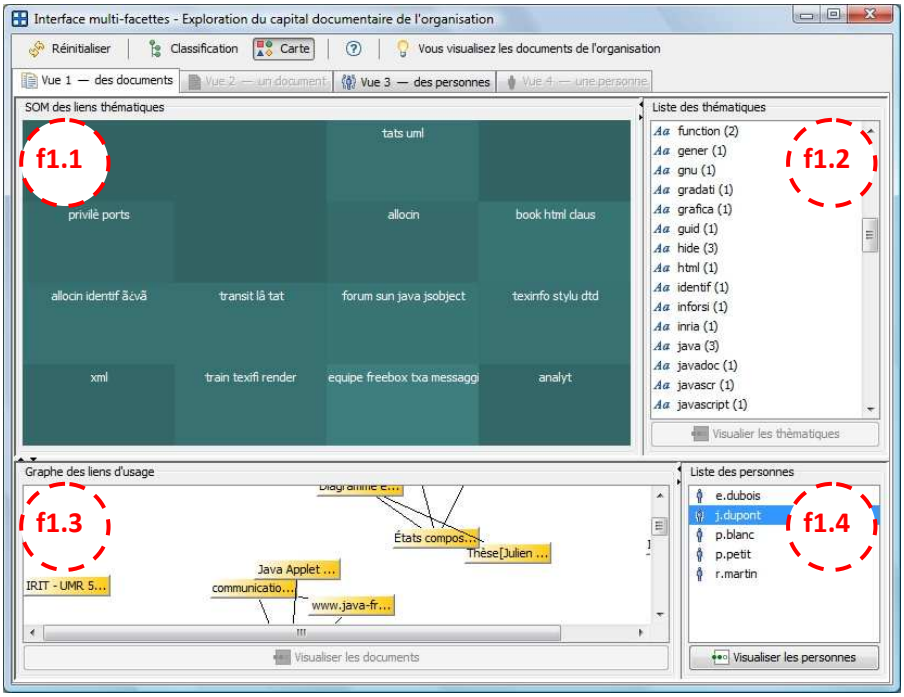


Figure 29 - Visualisation alternative dans la vue 1 pour une facette (f1.1)

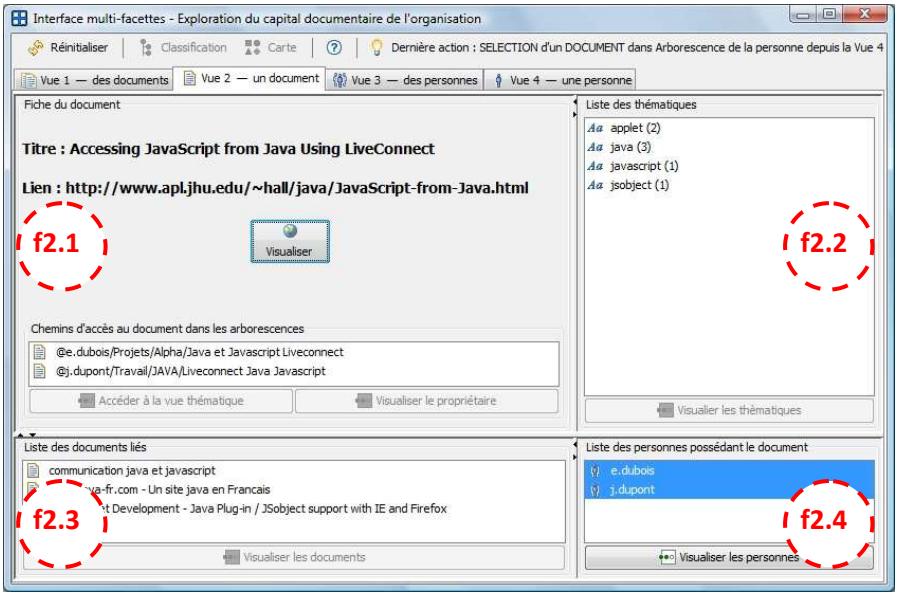


Figure 30 - Vue n°2 correspondant à un document spécifique

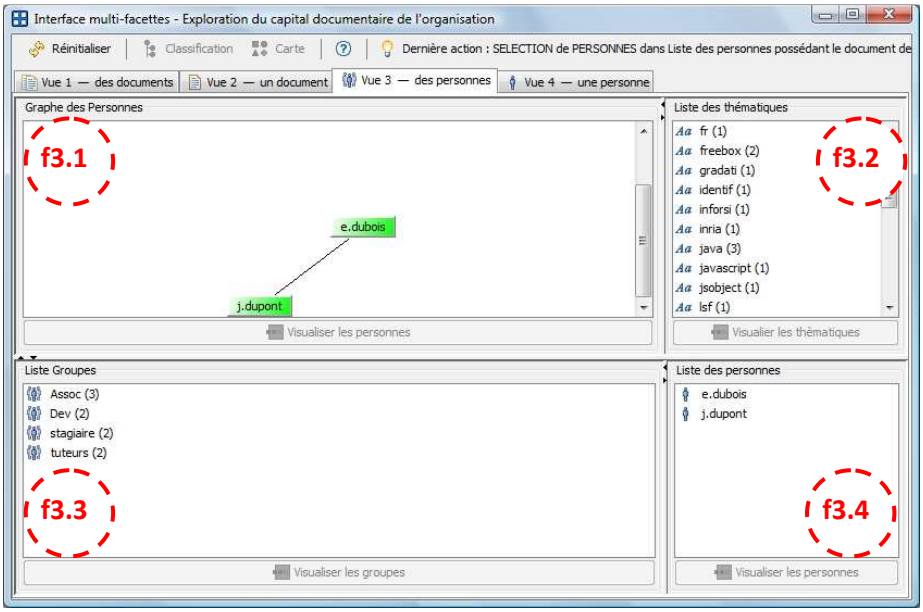


Figure 31 - Vue n°3 correspondant à un ensemble de personnes. Le « graphe des personnes » souligne les proximités thématiques entre usagers. La « liste groupes » précise les groupes auxquels l'usager est rattaché.

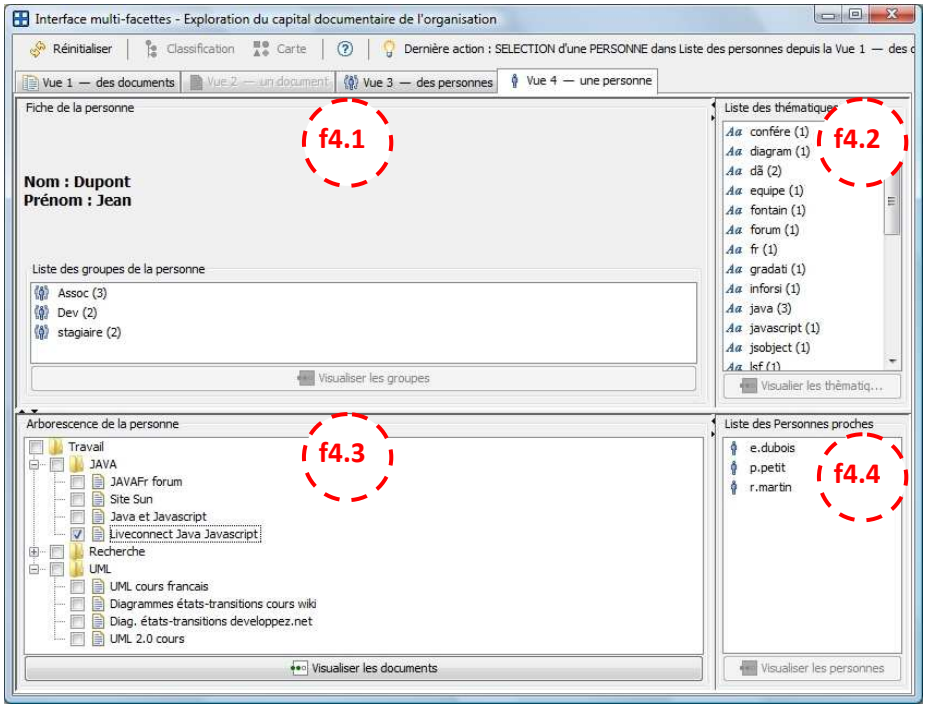


Figure 32 - Vue n°4 correspondant à une personne spécifique. Les personnes proches ont des intérêts thématiques proches de l'usager courant

Les copies d'écran précédentes illustrent les multiples possibilités offertes à l'usager lors de l'exploration des documents. Pour conclure sur la présentation de cette interface, la Figure 33 met en évidence les interactions possibles dans l'interface. Cette figure montre les relations entre les différentes vues en fonction du type de sélection faite par l'usager.

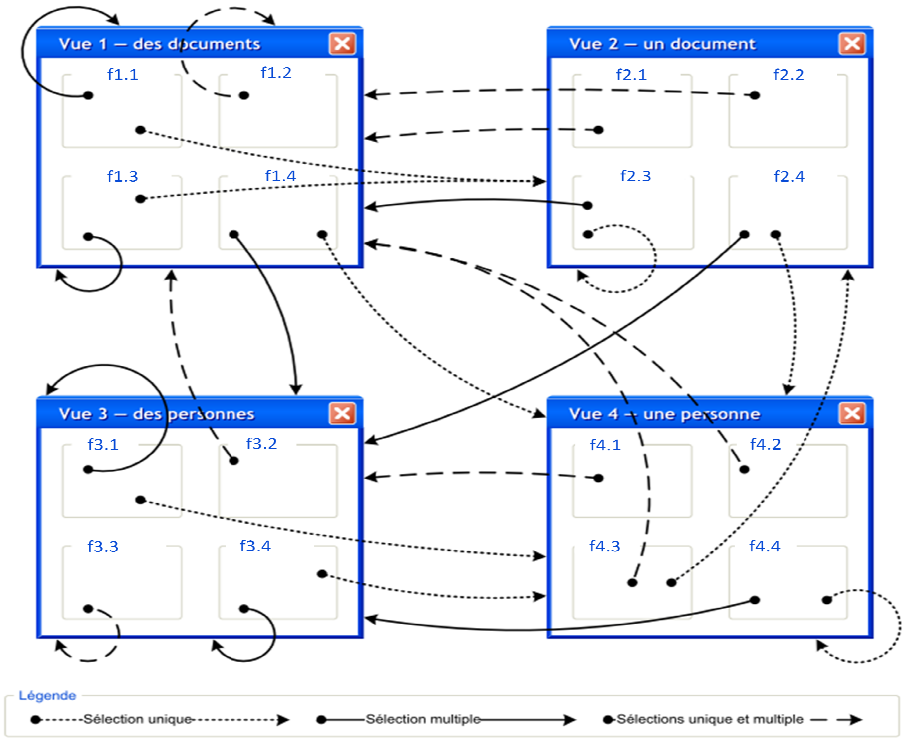


Figure 33 - Interaction sous-jacente à l'interface multi-facettes proposée

D.4.3.2 Plateforme de mesure de l'efficacité des interfaces de visualisation

Pour mesurer l'efficacité des interfaces de visualisation des résultats de recherche, nous avons implanté une plateforme complète (basée sur une architecture client/serveur) reprenant le protocole proposé. L'architecture proposée est présentée dans la Figure 34.

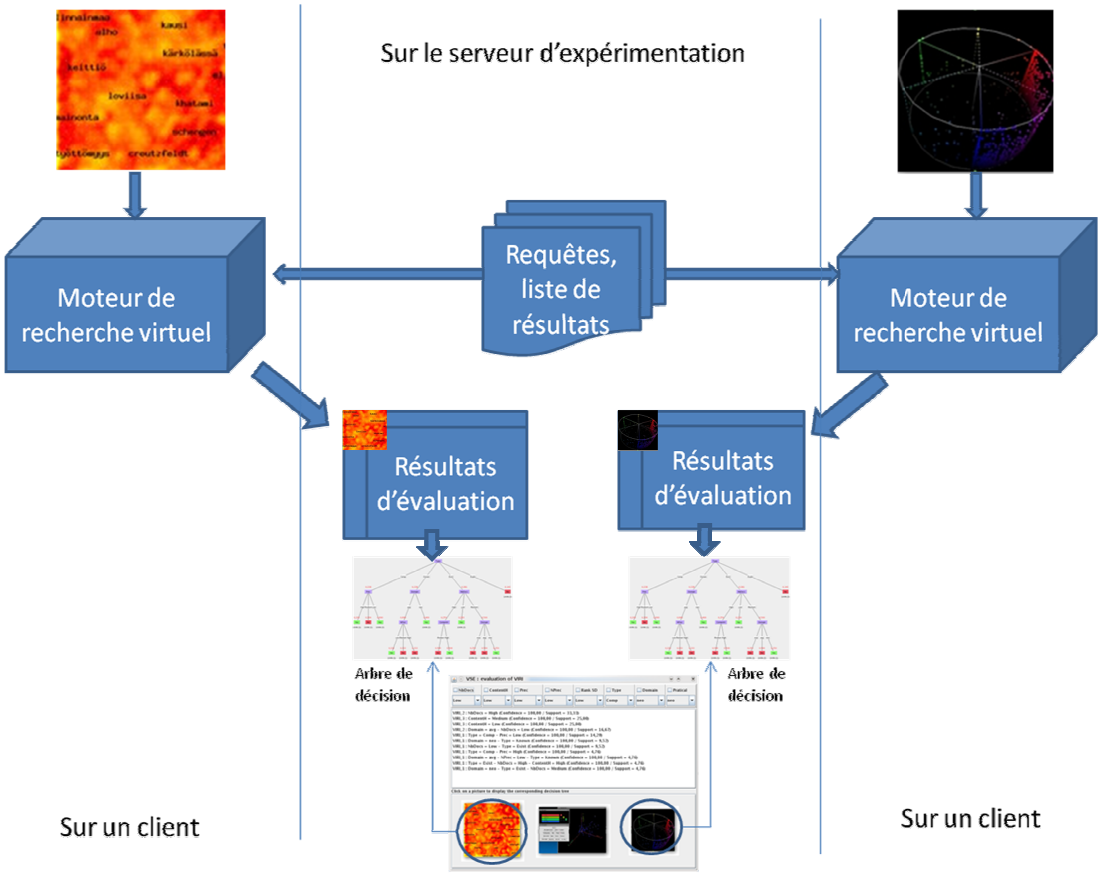


Figure 34 - Plateforme de mesure de l'efficacité des interfaces de visualisation pour la RI

Dans cette figure, la fenêtre située en dessous des arbres de décision a pour objectif de permettre d'explorer les résultats obtenus de façon globale. Ainsi cela permet à un usager de savoir, pour un contexte donné, quelles interfaces permettent d'obtenir généralement une meilleure efficacité (Figure 35). Pour ce faire, les règles menant à des succès sont extraites de l'ensemble des arbres de décisions du système. La fenêtre proposée présente initialement à l'usager toutes les règles ordonnées par leur valeur de *support* et de *confiance*. L'usager peut fixer ou non les valeurs des différents critères lui permettant ainsi de caractériser le contexte qui l'intéresse afin de filtrer les règles en conséquence. En bas de la fenêtre, les images des interfaces efficaces (c'est-à-dire pour lesquelles au moins une règle est présente dans la fenêtre) sont présentées. Un clic sur une image permet de visualiser l'arbre de décision correspondant à l'interface correspondante.

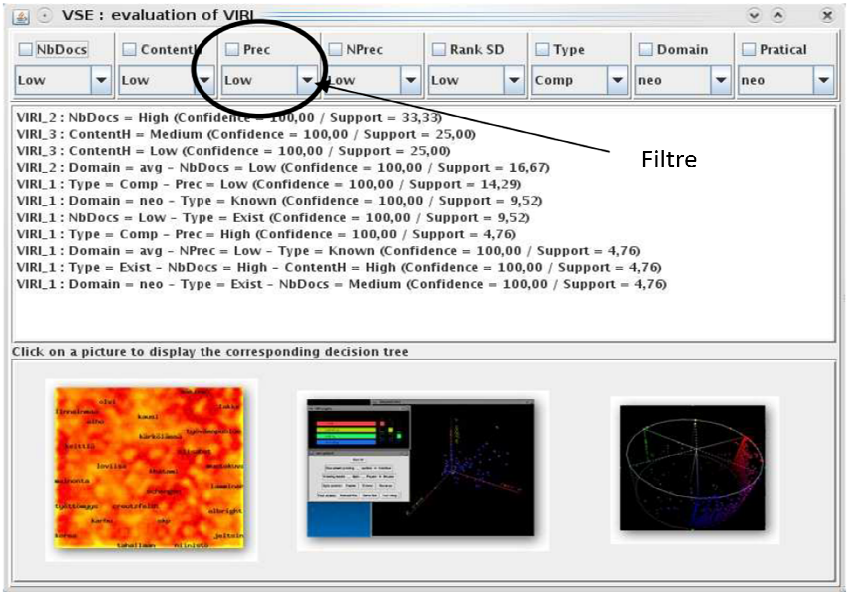


Figure 35 - Interface d'interrogation des interfaces de visualisation

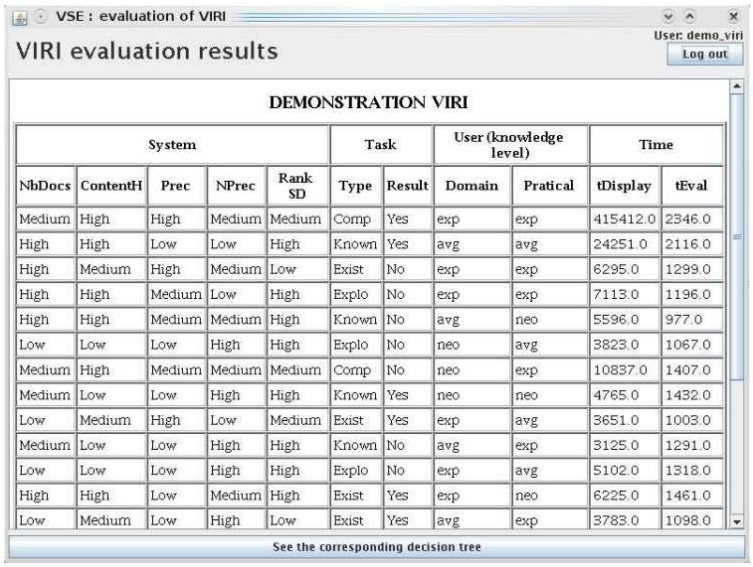
Le moteur de recherche virtuel, implanté en java, peut être exécuté sur n'importe quel système dès l'instant où un accès internet est disponible pour lui permettre de communiquer avec le serveur distant. Chaque interface communique avec le moteur soit par le biais de l'API (pour le moment uniquement implantée en Java) soit, par soucis de portabilité, au travers d'un protocole réseau permettant à n'importe quelle interface de visualisation développée dans n'importe quel langage de se connecter au moteur de recherche virtuel.

A tout moment, le concepteur d'une interface peut accéder aux résultats obtenus par son interface au travers du moteur de recherche virtuel (Figure 36) ainsi qu'à l'arbre de décision correspondant (Figure 37).

La génération des arbres de décisions ainsi que l'extraction des règles sont assurées par une bibliothèque de fonctions GINNet[20] développée par l'INRIA.

Une description plus complète de cette plateforme et de ses fonctionnalités est disponible dans (Port, 2009).

[20] http://ginnet.gforge.inria.fr



System					Task		User (knowledge level)		Time	
NbDocs	ContentH	Prec	NPrec	Rank SD	Type	Result	Domain	Pratical	tDisplay	tEval
Medium	High	High	Medium	Medium	Comp	Yes	exp	exp	415412.0	2346.0
High	High	Low	Low	High	Known	Yes	avg	avg	24251.0	2116.0
High	Medium	High	Medium	Low	Exist	No	exp	exp	6295.0	1299.0
High	High	Medium	Low	High	Explo	No	exp	exp	7113.0	1196.0
High	High	Medium	Medium	High	Known	No	avg	neo	5596.0	977.0
Low	Low	Low	High	High	Explo	No	neo	avg	3823.0	1067.0
Medium	High	Medium	Medium	Medium	Comp	No	neo	exp	10837.0	1407.0
Medium	Low	Low	High	High	Known	Yes	neo	neo	4765.0	1432.0
Low	Medium	High	Low	Medium	Exist	Yes	exp	avg	3651.0	1003.0
Medium	Low	Low	High	High	Known	No	avg	exp	3125.0	1291.0
Low	Low	Low	High	High	Explo	No	exp	avg	5102.0	1318.0
High	High	Low	Medium	High	Exist	Yes	exp	neo	6225.0	1461.0
Low	Medium	Low	High	Low	Exist	Yes	avg	exp	3783.0	1098.0

Figure 36 - Affichage des résultats d'expérimentation pour une interface spécifique

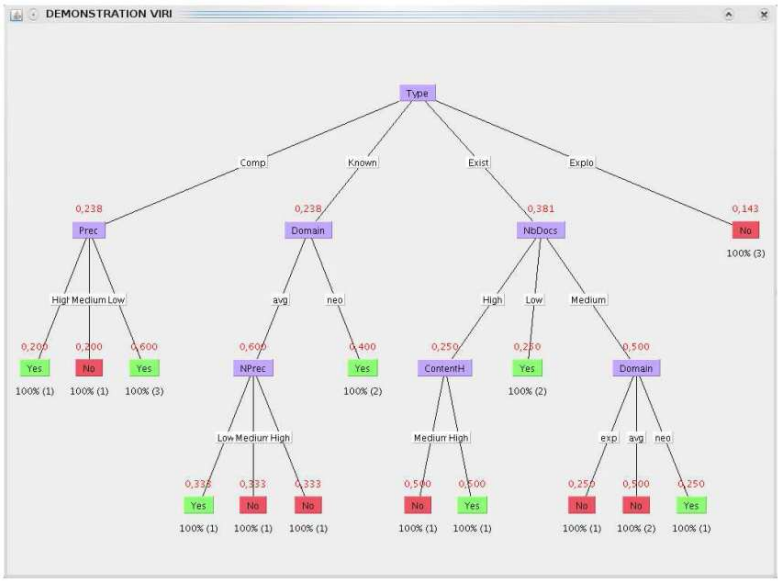


Figure 37 - Arbre de décision de l'interface ayant obtenu les résultats présentés dans la Figure 36

Pour conclure sur la présentation de cette plateforme de mesure de l'efficacité des interfaces de visualisation en RI, nous pouvons souligner le fait qu'elle :

- permet de mettre en évidence les **contextes dans lesquels une interface de visualisation est efficace** et à l'inverse identifier **pour un contexte donné les interfaces de visualisations efficaces**,
- est **extensible**. De nouvelles caractéristiques peuvent être ajoutées pour caractériser plus finement un contexte de recherche,
- est **portable**. Elle peut ainsi être utilisée sur différents systèmes d'exploitation,
- est **fonctionnelle** en l'état car elle est d'ores et déjà associée à 1000 listes de résultats (relatives à 50 requêtes TREC) caractérisant une multitude de contextes différents.

Les travaux précédents mettent en évidence les principales difficultés auxquelles peut être confronté un usager lors d'une recherche d'information. Un des éléments les plus importants que nous avons identifié est la portée faible des SRI, c'est-à-dire qu'aujourd'hui, en exagérant, le SRI n'est pas au service de l'usager. En effet, c'est finalement à l'usager qu'incombent les plus gros des traitements ; traitements qu'il réalise à la main. Cette situation peut s'expliquer par le fait que le système n'a que peu de moyens pour identifier les besoins (au travers d'une requête courte) et les objectifs réels de la recherche (que veut-il faire de ces informations ?). De nombreux travaux notamment autour des approches personnalisées et plus récemment autours des approches basées contexte (Tamine, 2008) visent à prendre en considération des éléments complémentaires et périphériques à la RI. Nous avons souhaité dans nos travaux explorer une piste parallèle : permettre à l'usager de contrôler sa recherche de bout en bout.

D.5 Donner le contrôle de la RI à l'usager

La personnalisation de la RI nécessite que le système puisse modéliser l'usager sur la base d'informations collectées. Cependant certaines dimensions, qui impactent une recherche, sont plus difficiles à mesurer : les intentions de l'usager (Tomasic, et al., 2007), le type de recherche ou encore des éléments plus subjectifs tels que la satisfaction de l'usager, la confiance qu'il a dans les résultats (Gwizdka, et al., 2009), son état affectif (Picard, 1997), etc.

Puisque le système n'a pas toutes les informations à sa disposition pour apprécier de façon précise l'usager et ses objectifs, la piste que nous explorons consiste à permettre à l'usager de construire sa RI à partir de briques de bases (métaphore des Lego™).

Cette réflexion est le résultat de l'analyse de l'activité de RI des chercheurs en génomique que nous avons réalisée dans le cadre du projet GenIE (Chevalier, et al., 2006) financé par l'Université Paul Sabatier. Une des conclusions de cette analyse est que les chercheurs devaient manipuler de grandes quantités d'informations provenant de recherches dans des sources hétérogènes (base de nucléotides, base de protéines, base d'articles scientifiques…) avec pour objectif l'identification de relations entre des objets biologiques. Cette tâche laborieuse ne peut pas être automatisée car une grande part d'expertise est nécessaire. Ce problème peut être résumé par la question : *qui d'autre que l'usager sait vraiment ce qu'il souhaite et sait ce qu'il veut en faire ?* Une conclusion assez proche peut être formulée dans d'autres contextes tels que les entrepôts de données (Ravat, et al., 2008).

Dans le même temps, une analyse empirique de la recherche d'information (Belbèze, et al., 2007)(Belbèze, et al., 2007) nous a, entre autre chose, permis de souligner que certaines caractéristiques humaines, différentes de celles présentées dans les chapitres précédents, telles que les déficiences (daltonisme ou encore problème de langue) n'étaient pas prises en compte par les outils de recherche.

Fort de cette situation, nous nous sommes naturellement orientés vers des plateformes reposant sur une architecture flexible à base de composants et plus précisément à base de Web services dans un premier temps. De là est née la plateforme nommée **Sissi** (Belbèze, et al., 2008). L'objectif est de permettre aux usagers de construire une recherche d'information ***complète*** (du choix de la source d'information, aux traitements réalisés sur la liste de résultats de recherche…) et ce sur la base des Web services disponibles.

Ces travaux font écho aux travaux de (Ferreira, et al., 2005), (Meyer zu Eissen, et al., 2007). Ces approches sont assez proches de nos motivations sauf qu'elles se focalisent uniquement sur le processus même de RI (structure d'un outil de recherche). L'utilisation de ces approches n'est donc pas généralisable car une bonne expertise de la RI est requise. Dans notre cas, nous nous intéressons à la chaîne complète de traitement liée à la recherche d'information et souhaitons qu'elle soit accessible à tous les publics. Ainsi, nous privilégions le fait que les outils de recherche actuels (les moteurs de recherche Google ou Bing par exemple) ne sont utilisés que comme des boîtes noires. Aucune expertise n'est donc demandée à l'utilisateur final.

Plus récemment (Juin, 2010), nous avons implanté une seconde plateforme à base de composants au sein de laquelle on peut ajouter « à chaud » des composants logiciels Java. Les composants peuvent être recherchés soit par catégorie soit par tags (Figure 38, partie de gauche). L'usager doit positionner et relier les composants nécessaires au traitement souhaité. Certaines caractéristiques des composants peuvent être configurées (exemple : la requête pour le composant du moteur de recherche Bing). La Figure 38 présente une chaîne de traitement « de base » puisqu'il s'agit d'interroger le moteur de recherche Bing et d'afficher les résultats sous la forme d'une liste de résultats (Figure 39).

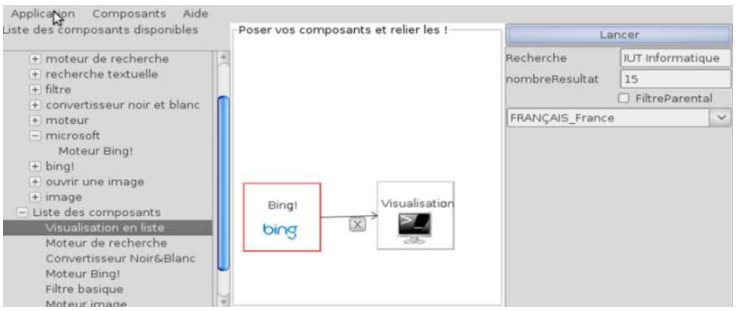


Figure 38 - Plateforme à base de composants pour la recherche d'information

Figure 39 - Liste de résultats obtenue par l'exécution de la chaîne de traitement présentée dans la Figure 38

L'intérêt de cette plateforme est qu'elle permet d'obtenir des traitements complets et complexes tel que celui présenté dans la Figure 40. L'objectif de ce traitement est le suivant : *Le moteur de recherche Bing (recherche textuelle) et Bing-Image (recherche d'images) sont interrogés avec la même requête « allociné l'homme sans passé » (c'est-à-dire une recherche liée au film « L'homme sans passé »). Toutes les images provenant du moteur Bing-Image*

sont traitées et converties en noir et blanc (pour limiter les dyschromatopsies[21]*) afin que les résultats puissent être exploités par l'usager. Tous les résultats textuels provenant de Bing sont filtrés (pour ne conserver que les documents retrouvés ayant le mot « passé » dans le titre). L'avant dernier composant « Fusionner Image et Texte » permet de rassembler les textes et les images provenant de la même URL avant d'être renvoyés à l'usager sous la forme d'une liste de résultats.*

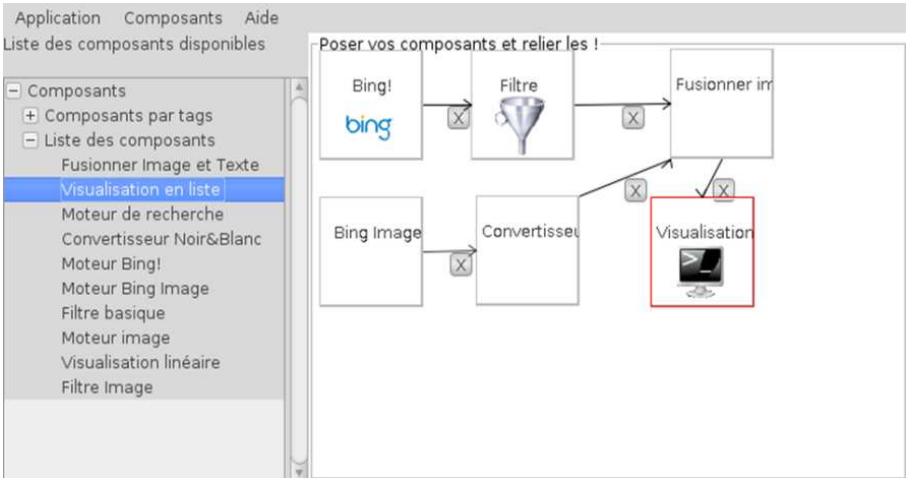


Figure 40 - Exemple de traitement avec transformation d'images

En conclusion, l'architecture de RI à base de composants que nous proposons :

- est **extensible** et **ouverte**. Quiconque peut insérer un nouveau composant (une API est proposée) et le mettre à disposition de la communauté,
- est **accessible**. Elle permet à tous les usagers d'exprimer facilement au travers des composants existants ses besoins en termes de traitement. L'usager obtient donc plus aisément le résultat de recherche qui l'intéresse.

D.6 Bilan et résultats

Dans ce chapitre, nous nous sommes intéressés à l'activité de RI en nous focalisant sur les problématiques essentiellement liées à l'usager. Ainsi, les orientations de nos travaux sont relatives à :

- la formulation des besoins,
- l'impact de la visualisation dans le cadre de la RI,
- le contrôle de la RI par l'usager.

D.6.1 La formulation des besoins

Dans cet axe de travail, nous avons proposé une approche originale de désambigüisation des requêtes selon une approche par médiation. Cette approche repose sur la construction d'agrégats de termes issus des requêtes précédemment soumises à un moteur de recherche. Ainsi, le système propose à l'usager différents contextes d'usage des termes de sa requête pour qu'il puisse lui-même comprendre comment modifier sa requête.

[21] Déficiences de la vision colorée

La force de notre proposition réside dans le fait que :

- l'usager reste **maître de la situation**. Le système ne se substitue pas à l'usager, il agit seulement comme assistant en lui indiquant les différents contextes d'usage des termes de sa requête,
- les contextes d'usage des termes sont **dynamiques**. Ces contextes évoluent dans le temps au même rythme que les termes utilisés dans les requêtes évoluent.

D.6.2 La visualisation en RI

Les travaux concernant cet axe de recherche soulignent l'intérêt des techniques de visualisation dans le cadre de la RI. L'originalité des propositions porte sur :

- une interface multi-facettes permettant de visualiser les résultats de RI dont les principales forces reposent sur :
  - la **synchronisation** entre les vues permettant à l'usager de se focaliser rapidement sur les informations répondant à ses besoins,
  - des **facettes complémentaires** permettant d'apprécier un objet globalement selon différents points de vue,
  - la possibilité d'être **étendue**. Elle peut être déclinée dans différents contextes applicatifs.
- la **mesure de l'efficacité des interfaces de visualisations** des résultats de recherche dont les principales forces reposent sur :
  - la définition d'un **contexte de recherche d'information** <Usager, Système, Tâche> extensible dans lequel chacune des dimensions est caractérisé par un ensemble de caractéristiques connues dans le monde de la RI,
  - un **protocole d'expérimentation complet**. Il permet en effet de connaître, pour une interface, les contextes de recherche d'information dans lesquels elle est efficace,
  - une plateforme **complète** et **extensible** reprenant le protocole proposé. Cette plateforme est également **portable**.

D.6.3 Donner le contrôle de la RI à l'usager

L'analyse du comportement des usagers face à la RI (notamment dans le projet GenIE) nous a conduit à nous interroger sur la possibilité de donner une maîtrise totale à l'usager lui permettant de construire ce qu'est pour lui une RI. Nous avons ainsi proposé l'utilisation d'une plateforme à base de composants permettant de structurer les différents traitement qui peuvent être utiles pour l'usager en amont ou en aval de la RI.

L'originalité de nos travaux se retrouvent dans le fait qu'ils reposent sur une plateforme qui :

- tient compte des **motivations, objectifs, handicaps** des usagers puisqu'ils traduisent eux-mêmes ces éléments en traitements,
- permet à l'usager de **contrôler la chaîne de RI** dans sa globalité (en intégrant les composants souhaités dans la chaîne de traitement),
- est **extensible** par de nouveaux composants que quiconque peut développer sur la base de l'API fournie,

- est **portable** sur différents systèmes. L'implantation en Java qui en a été faite favorise cette portabilité.

En conclusion de ce chapitre, nous pouvons souligner les trois principales leçons que nous pouvons tirer de nos travaux :

- l'usager doit pouvoir garder le contrôle de son activité. Dans ce cadre, il est intéressant de privilégier les approches par médiation,
- les interfaces de visualisation jouent un rôle important tant en amont (visualisation des requêtes) qu'en aval de la RI (visualisation des résultats de recherche),
- il est important de donner à l'usager toutes les informations nécessaires pour qu'il puisse apprécier globalement et de façon la plus complète possible les objets qui lui sont présentés (ainsi que les relations que ces objets entretiennent entre eux) : approches multi-facettes.

D.7 Valorisation et support à la recherche[22]

Ces travaux servent de support à une thèse et ont donné lieu à 12 publications.

Encadrement	Publications				Projets associés
	RI	RN	CI/WI	CN	
1 thèse (C. Belbèze) et 1 M2 (R. Amara)		2	4	6	*GenIE*

Les propositions relatives à l'aide à la formulation des besoins (désambigüisation des requêtes et système de recherche à base de composants) suivent la logique des résultats obtenus dans le projet GenIE et sont essentiellement développées dans la thèse de Christian Belbèze.

Les travaux relatifs à la visualisation multi-facettes de résultats de recherche font suite aux publications telles que (Chevalier, et al., 2005) et ont été développées dans la thèse de Guillaume Cabanac.

La plateforme de mesure de l'efficacité des interfaces de visualisation des résultats de recherche est le fruit d'un travail mené en collaboration avec notamment Nicolas Bonnel, Bernard Dousset et Gilles Hubert.

[22] Plus de détails concernant les thèses et Master 2 encadrés sont fournis en section F.1 (p. 103).

D.8 Perspectives

Les travaux que nous avons menés concernant l'amélioration des SRI peuvent encore évoluer pour donner à l'usager la possibilité de réaliser sa RI dans de meilleures conditions. Ainsi plusieurs perspectives dans ce cadre peuvent être envisagées.

D.8.1 Perspectives concernant la formulation des besoins

Certains usagers, ne connaissant pas bien le domaine de recherche et l'outil qu'ils utilisent (la base d'indexation notamment) peuvent avoir du mal à formuler leurs requêtes. L'approche de désambigüisation basée sur une médiation que nous proposons permet à l'usager d'identifier les « contextes » dans lesquels les mots qu'il utilise pour sa requête se retrouvent. Grâce à cela l'usager peut adapter sa requête en fonction des contextes qui l'intéressent. Cependant, cette approche repose sur l'utilisation des logs de requêtes pour identifier ces contextes. Cela signifie que les contextes proposés par le système évoluent en fonction du temps au même rythme que la langue elle-même. Cet aspect est à la fois une force mais également une faiblesse. En effet, les requêtes posées par les usagers peuvent ne pas produire de résultats satisfaisants, l'outil de recherche ayant un impact important dans ce cadre car c'est lui qui réalise l'appariement entre les documents et la requête. Ainsi une perspective possible pourrait consister à **donner une dimension contextuelle à notre approche** en utilisant conjointement les mots contenus dans le fonds documentaire et des mots issus des requêtes pour construire les agrégats. Ainsi, certains mots non présents dans la base documentaire pourraient être pondérés voire supprimés des contextes construits. Cette démarche permettrait de mieux contextualiser les requêtes par rapport au moteur de recherche. L'approche pourrait ainsi se limiter à la présentation des termes pour lesquels l'outil de recherche possède des résultats et faciliterait dans le même temps la médiation.

Une deuxième perspective dans ce cadre concerne le **passage à l'échelle**. Pour obtenir des performances adéquates particulièrement en temps de réponse, surtout dans le cadre du Web, nous devons travailler à une optimisation des traitements liés au graphe. Les collaborations que nous entretenons avec des mathématiciens experts de ce domaine vont nous permettre vraisemblablement d'atteindre cet objectif.

D.8.2 Perspectives concernant la visualisation en RI

A l'autre bout de la chaîne, les interfaces de visualisation ou d'exploration pour la RI ne sont qu'à un premier stade et différentes perspectives ont été identifiées.

L'interface multi-facettes que nous avons proposée est actuellement fonctionnelle. Une première perspective serait de la rendre **générique**. En effet, les interactions possibles entre les différentes facettes sont pour le moment figées par le programmeur. Une perspective serait de permettre au concepteur (ou mieux à l'usager) de configurer l'interface pour ajouter des interfaces ainsi que des interactions entre les facettes sans avoir à reconstruire l'application. XML dans ce cadre devrait pouvoir jouer pleinement son rôle.

La plateforme de mesure de l'efficacité des interfaces est fonctionnelle. La première perspective concerne le lancement d'une phase d'expérimentation et de **mesure effective de l'efficacité de différentes interfaces**. Pour le moment une seule interface (sous la forme de liste de résultats) est proposée avec la plateforme. Il est donc nécessaire de motiver des concepteurs d'interfaces pour qu'ils adaptent leurs interfaces de sorte à ce qu'elles puissent être évaluées sur notre plateforme. Ensuite le recrutement d'évaluateurs est envisagé par le biais d'appels sur les listes de diffusion. L'application étant facilement exécutable au travers du Web, cela permet de toucher potentiellement un plus grand nombre de personnes.

Une seconde perspective concerne la génération de **jeux d'essai** sur la base desquels la mesure d'efficacité est réalisée. Il faut développer des jeux d'essai correspondant aux différentes tâches que nous avons identifiées (D.4.2.1). Ce travail ne peut être réalisé que sur la base de coopérations entre chercheurs car la tâche est très importante. C'est pour cela que pour le moment la plateforme ne permet d'évaluer qu'une seule tâche : la recherche d'une information quelconque.

Enfin, une dernière perspective concerne la **modélisation de l'efficacité** que nous avons proposée. De nouveaux critères supplémentaires pourraient être proposés pour affiner la mesure de l'efficacité. Par ailleurs, la collecte d'informations relatives à l'utilisation de l'interface testée permettrait une analyse conjointe des résultats de recherche renvoyés par la personne testeur et du log d'interaction. Cette analyse aiderait les concepteurs des interfaces à comprendre plus précisément les limites ou les atouts d'une interface de visualisation dans un contexte particulier.

A plus long terme, une perspective plus globale relative à nos travaux sur les interfaces serait de les faire **converger (approche multi-facette et évaluation de l'efficacité)**. En effet, il serait intéressant sur la base des résultats que nous obtenons avec la plateforme de mesure de l'efficacité des interfaces de les intégrer dans l'interface multi-facettes pour que le système puisse, de façon adaptative, guider l'usager vers les visualisations disponibles qui, dans son contexte, lui permettrait d'identifier rapidement les documents pertinents.

D.8.3 Perspectives concernant la prise de contrôle de la RI par l'usager

Concernant le système à base de composants pour la RI, une perspective à court terme consiste à le mettre en ligne afin de vérifier l'intérêt qu'il suscite auprès des usagers réels et d'en faire une évaluation qualitative.

A plus long terme une perspective intéressante est de proposer des **outils d'analyse** des résultats de recherche directement sous forme de composants. En parallèle à cela, une évolution passerait par l'introduction dans un tel système d'un **processus d'apprentissage** permettant d'aider l'usager dans le choix des composants à utiliser.

D.9 Perspectives concernant l'utilisation des SRI

Lors de l'utilisation d'un SRI, on se rend compte que l'activité de l'usager est conséquente et souvent peu évidente (exemple : convertir des besoins en requête). Une perspective concernant les systèmes de recherche d'information serait d'intégrer plus d'outils permettant à l'usager d'explorer et rechercher dans les fonds documentaires d'intérêt et ce de façon transparente. En effet, avec le grand nombre de moteurs disponibles, avec l'émergence des systèmes de recherche « personnels » (permettant de rechercher des documents sur notre ordinateur personnel), avec la maturité des outils de fusion de résultats de recherche, des outils de recherche multi-facette et des interfaces de visualisation, une solution pourrait être trouvée dans la définition d'un **méta-moteur de recherche et d'analyse** d'informations.

E. MODÈLES POUR UN ACCÈS PERSONNALISÉ À L'INFORMATION

E.1 Introduction

Les systèmes adaptatifs en RI reposent sur des modèles représentant en particulier l'usager. Il est important de noter que l'usager peut être amené à utiliser plusieurs applications notamment sur le Web pour obtenir des informations répondant à ses besoins. Du point de vue de la modélisation de l'usager cela implique que chaque application possède une « portion » du modèle de l'usager réalisant sa recherche. A aucun moment les différentes applications ne possèdent une vision complète du modèle de l'usager, ce qui tend à minimiser la qualité de l'adaptation proposée par chaque application.

Par ailleurs, dans un souci d'amélioration continue de la RI, ces systèmes adaptatifs ont intégré une dimension plus large du **contexte** pour compléter la vision qu'ils possèdent de l'activité de l'usager.

Ce chapitre a pour objectif de présenter ces deux problématiques :

- la **modélisation de l'usager** dans un contexte de RI,
- la **modélisation du contexte** pour la RI.

Pour chacune de ces problématiques, une proposition de solution est apportée dans les sections E.3 et E.4.

E.1.1 Modélisation de l'usager dans un contexte multi-applications

La modélisation de l'usager repose sur une grande variété de modèles issus de la littérature (section B.2.2.2). Dans le cadre de la RI, l'usager utilise conjointement différents systèmes ayant tous pour objectif de l'aider dans sa démarche. Il s'agit donc d'un *contexte multi-applications*. Cet état de fait pose le problème de gestion des profils usagers et l'obtention d'un modèle complet[23] de l'usager permettant d'accroître la qualité des adaptations de chaque application. Par exemple, lors d'une recherche, un moteur de recherche adaptatif n'exploite que les centres d'intérêt d'un usager qui sont construits sur la base des informations qu'il détient. Il **n'exploite donc pas les centres d'intérêts modélisés par un autre moteur de recherche qui vient d'être utilisé par l'usager.**

Pour illustrer ce problème, nous présentons les différentes relations possibles entre les systèmes pour obtenir une vision complète de l'usager dans un contexte multi-applications comme c'est le cas dans la RI.

[23] *Complet* dans ce cadre ne signifie pas exhaustif. Un modèle complet correspond à un modèle incluant un maximum de caractéristiques de l'usager dans la limite de ce qu'il est possible d'identifier par les systèmes et de ce qui est nécessaire pour les traitements.

Tout d'abord la situation la plus répandue consiste à avoir un ensemble de systèmes totalement indépendants. Du point de vue de la modélisation usager cela implique que chaque application modélise d'une manière qui lui est propre (Figure 41).

Ceci induit (du point de vue de la modélisation) que :

- ☑ Chaque application reste maîtresse du modèle utilisé. Elle possède sa propre vision de l'usager (liberté en terme de structure, de contenu du modèle usager),
- ☒ Aucun des systèmes n'a de modèle complet.

Pour réaliser sa recherche, l'usager a la possibilité d'utiliser n'importe quelle application disponible.

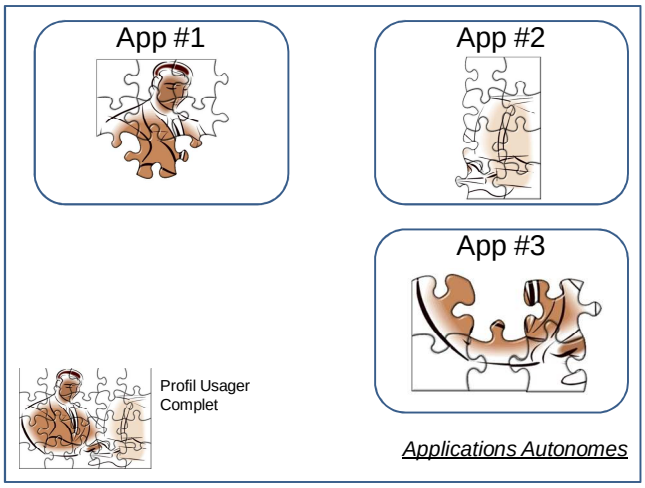


Figure 41 - Cas le plus courant : les applications sont indépendantes

Dans l'objectif de proposer à chaque application un modèle plus complet de l'usager et limiter ainsi la redondance des informations liées au profil, une alternative peut être trouvée dans le développement d'une application unique intégrant les différentes fonctionnalités proposées initialement dans les autres applications (Figure 42). L'implantation de l'architecture pour la fédération des activités documentaires que nous avons proposée (section C.5) repose sur ce principe.

Ceci induit (du point de vue de la modélisation) que :

- ☑ Toutes les applications partagent le même modèle usager ce qui permet d'avoir une vision complète et à jour de l'usager,
- ☑☒ Le modèle initial doit être construit sur la base d'un consensus ce qui peut ne pas être simple dans le cas où un grand nombre d'applications sont intégrées. Ceci entraîne donc une perte de liberté dans la modélisation.

Si l'on souhaite obtenir une vision complète du modèle usager, la principale limite de cette approche n'est pas liée à la modélisation de l'usager mais bel et bien à l'usage de ce type d'applications. En effet, le choix de l'usager est limité à une unique application (impliquant un changement d'habitudes) alors qu'il pourrait préférer l'usage d'applications distinctes qui lui correspondent mieux.

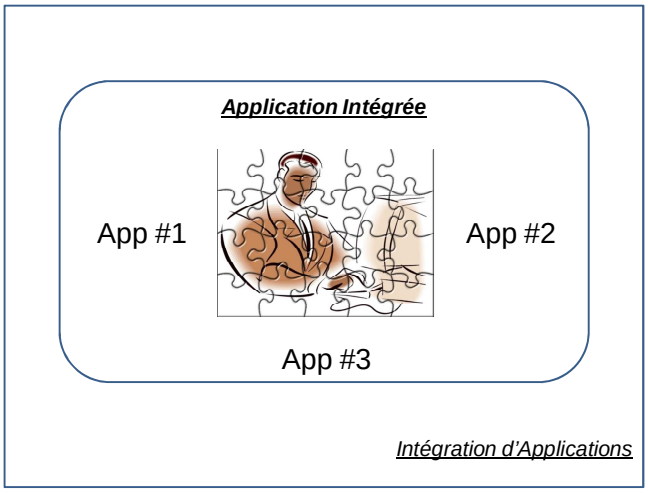


Figure 42 - Application Intégrée

Une troisième possibilité, qui repose également sur l'intégration, propose l'intégration uniquement des modèles usagers au travers par exemple d'un serveur de modèles (Figure 43). Différentes propositions dans ce cadre ont été faites dans la littérature et notamment les travaux relatifs à *UMS* (Kobsa, et al., 2006) et *Personis* (Kay, et al., 2002). Une analyse et évaluation des serveurs de modèles pour la personnalisation sur le Web a été menée dans (Fink, et al., 2000).

Ceci induit (du point de vue de la modélisation) les mêmes éléments que pour l'application intégrée c'est-à-dire que :

- ☑ Toutes les applications partagent le même modèle usager ce qui permet d'avoir une vision complète et à jour de l'usager (→ pas de redondance),
- ☑☒ Le modèle initial doit être construit sur la base d'un consensus ce qui peut ne pas être simple dans le cas où un grand nombre d'applications sont intégrées. Ceci entraîne donc une perte de liberté dans la modélisation.

Si l'on souhaite obtenir une vision complète du modèle usager, l'usager reste libre dans le choix des applications qu'il souhaite utiliser dès l'instant où celles-ci utilisent le serveur de profils.

Enfin, une dernière possibilité que nous avons identifiée est l'interopérabilité. L'interopérabilité entre les applications leur permet d'échanger des informations concernant les usagers (Figure 44).

Ceci induit (du point de vue de la modélisation) les mêmes éléments que pour l'application intégrée, c'est-à-dire :

- ☑ Toutes les applications ont la liberté de modéliser l'usager comme elles le souhaitent,
- ☑ Les applications possèdent une vision complète de l'usager,
- ☑☒ Il y a des risques importants de redondance des informations. Lors de l'échange il peut y avoir des incohérences entre les informations détenues par une application et les informations reçues.

E. Modèles pour un accès personnalisé à l'Information

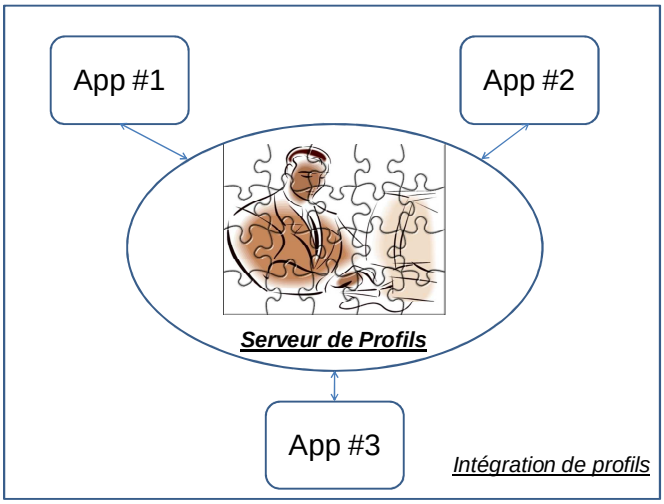


Figure 43 - Intégration des applications par les données : serveur de profils

Si l'on souhaite obtenir une vision complète du modèle usager, les applications peuvent être choisies librement par l'usager dès l'instant où elles supportent cette interopérabilité.

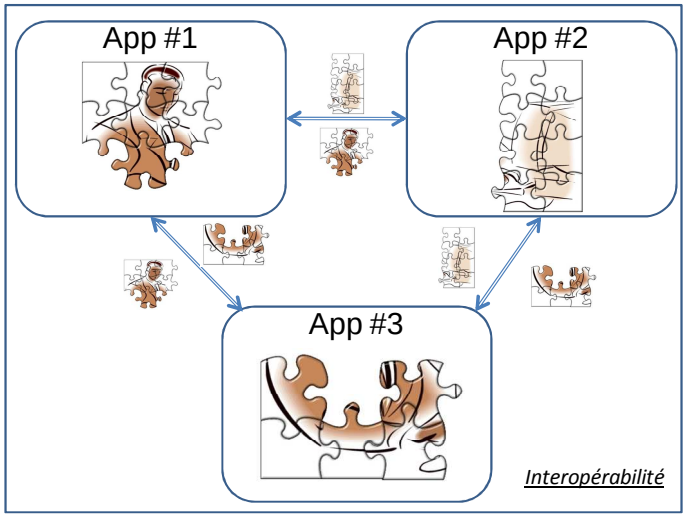


Figure 44 - Architecture flexible : interopérabilité des applications

Pour conclure, nous pouvons résumer les différentes possibilités par :

	Liberté de modélisation de chaque application	Vision complète de l'usager	Incohérences du modèle complet possibles	L'usager est libre de choisir les applications qu'il souhaite utiliser
Indépendance des applications	**+**	**-**	**n/a**	**+**
Application intégrée	**-**	**+**	**=**	**-**
Serveur de modèles	**-**	**+**	**=**	**=**
Interopérabilité	**+**	**+**	**=**	**=**

Tableau 6 - Synthèse des différentes possibilités d'obtention d'un modèle usager complet dans un contexte multi-applications.
(Légende : + impact positif important ; - impact négatif important ; = impact modéré ; n/a : non applicable)

On peut constater au travers de cette synthèse que si l'on prend en considération principalement la vision complète de l'usager en respectant la liberté pour un usager de choisir les applications qu'il souhaite, les approches basées sur les serveurs de modèles ainsi que l'interopérabilité sont les plus adaptées.

Cependant utiliser un serveur de modèles ne permet pas aux applications de modéliser l'usager comme elles le souhaitent. Or, notamment sur le Web, les informations détenues sur un usager représentent une mine d'or (commerciale) pour les applications qui les possèdent. Ainsi, en limitant cette liberté on limite dans le même temps l'intérêt qu'ont les concepteurs de systèmes à participer à la construction d'un modèle complet de l'usager. Nous pensons donc que l'interopérabilité des systèmes est une solution plus séduisante car le système reste maître de ce qu'il échange avec les autres applications sur un principe donnant-donnant et qu'il est libre de modéliser l'usager comme bon lui semble.

E.1.2 Modélisation du contexte pour la RI

La modélisation usagers est un élément central des systèmes adaptatifs. Cependant, se concentrer uniquement sur l'usager limite les possibilités d'adaptation. Ainsi, pour obtenir une vision élargie de l'usager, les systèmes adaptatifs en RI intègrent le contexte.

Dans le Petit Larousse Illustré de 1996 p.265, le *contexte* est défini par « 2. Circonstance, situation globale où se situe un évènement ». Il a été également défini dans différents domaines tels que la psychologie, l'informatique ubiquitaire, la recherche d'information etc. C'est une notion difficile à cerner tant les définitions peuvent varier d'un contexte applicatif à l'autre. (Bazire, et al., 2005) soulignent d'ailleurs « qu'il est difficile d'identifier une définition pertinente satisfaisante dans toutes les disciplines ».

Dans le domaine plus spécifique de la RI, le contexte est considéré comme multidimensionnel (les dimensions pouvant varier entre les auteurs), (Cool, et al., 2002), (Johnson, 2003), (Ingwersen, et al., 2005), (Tamine, 2008). On peut identifier un cœur commun aux définitions proposées : l'environnement et les dimensions humaines. Cependant, il n'y a pas de consensus concernant le détail de ces dimensions (Brusilovsky, et al., 2007).

A l'issue de l'étude des travaux autour du contexte, nous pouvons conclure que :

- le contexte est une notion **« spécialisée »** : on ne peut pas décrire un contexte de façon globale mais dans le cas d'un objet (un objet pouvant être un usager, une tâche de recherche, un système…),
- un contexte peut avoir un **impact** sur l'objet considéré. Ainsi par exemple, le contexte d'un système peut impacter le fonctionnement du système,
- cette notion est **récursive** : Un objet contextualisé (situé dans un contexte) peut à son tour être un élément du contexte d'un autre objet.

D'après ces conclusions, nous pouvons définir la notion de contexte :

Définition. Le contexte d'un objet représente tous les éléments pouvant avoir un impact sur cet objet.

Ainsi, considérant l'objet « recherche d'information », le contexte peut être composé des caractéristiques de l'usager, de l'environnement de recherche, etc….

En étudiant la notion de contexte, un élément important est apparu : la relation entre la tâche métier et les tâches informationnelles. En effet, (Byström, et al., 2005) par exemple soulignent le fait que les activités de RI sont des sous-tâches des tâches de travail (tâche métier). Ainsi, du point de vue de l'activité de l'usager, la RI n'est pas une fin en soi et ne participe qu'à la réalisation d'une tâche liée à son activité de travail. La relation entre la tâche métier et la RI a également été soulignée dans (Li, et al., 2008) qui démontrent que le comportement interactif en RI dépend de la tâche métier. D'après notre définition, la tâche métier peut donc être intégrée au contexte de la RI.

E.2 Enjeux & Objectifs de nos travaux

Afin d'améliorer les systèmes adaptatifs en RI, nos propositions reposent sur les conclusions des sections précédentes et ont pour objectif de :

- dans un contexte multi-applications et afin de proposer aux différentes applications utilisées par l'usager lors d'une recherche, fournir une **architecture interopérable de profils**. Cette architecture repose essentiellement sur un modèle de profils hiérarchiques et un modèle pivot. Ce modèle pivot se situe au niveau sémantique et permet d'assurer l'échange d'information entre les différentes applications utilisant des modèles de profils hétérogènes,
- donner une vision plus large de l'usager aux applications. Nous avons abordé l'implantation du contexte en RI. En réponse aux conclusions tirées de l'état de l'art, nous définissons une **architecture de gestion de contexte** qui permet à un système d'obtenir une vision contextualisée des objets qu'il manipule.

Les objectifs de chacune de ces orientations sont complémentaires dans le sens où un profil usager peut être manipulé par le gestionnaire de contexte. En résultat, nous proposons au travers de ces deux architectures un support à l'adaptation des systèmes pour la RI. Les sections suivantes présentent les travaux poursuivis dans chacune de ces deux orientations.

E.3 Modèle de profils flexible et interopérable

Dans le cadre de la modélisation usager, l'enjeu d'obtenir une vision complète de l'usager pour un ensemble d'applications tout en préservant l'indépendance apparaît comme essentielle.

Pour cela, nous proposons une architecture de profil flexible et interopérable. Cette architecture repose sur des profils hiérarchiques et est particulièrement adaptée pour supporter les types de variables basés sur le modèle attribut-valeur ou vectoriel. Ce choix se justifie par l'importance de ces modèles dans les travaux existants de RI.

L'architecture repose sur deux éléments principaux (Chevalier, et al., 2007) :

- un modèle de ressources incluant un niveau sémantique servant de base à l'interopérabilité,

- une démarche d'identification des appariements possibles entre deux profils. Ce point est la première étape avant de pouvoir échanger des informations entre applications.

E.3.1 Modèle de ressources

Le modèle de profils proposé (Chevalier, et al., 2004), (Chevalier, et al., 2005b), (Chevalier, et al., 2005c) est présenté dans la Figure 45. Ce modèle est divisé en quatre niveaux :

- la **structure logique** du profil. Elle permet de décrire hiérarchiquement les différentes variables (nom, centre d'intérêts…) composant un profil. Nous pouvons souligner que la structure logique d'un profil peut être une composition de structures logiques d'autres profils. Ceci permet par exemple de construire indépendamment les différentes dimensions des profils pour ensuite les composer,

- la **sémantique associée à la structure logique** : elle permet de décrire la sémantique des éléments de la structure logique. Un profil permet de décrire une *Ressource* (usager, imprimante…). Le *LienRessources* permet d'établir les relations qui peuvent exister entre les différentes ressources. Chaque attribut sera également décrit de la même manière sur la base d'une description conceptuelle. Par exemple, l'attribut « centres d'intérêt » pourra être associé au concept « thématique »,

- le **contenu** du profil : le contenu correspond à la valeur de chaque *attributFeuille* d'un profil. Un *ElementDeContenu* correspond à une valeur de l'attribut à laquelle est associé un poids réel permettant ainsi de supporter le modèle vectoriel. Par exemple, un *attributFeuille* nommé « Centres d'intérêt » peut être associé à deux *ElementDeContenu* ("informatique", 0.8) et ("yoga", 0.4) => *centres d'intérêts = {("informatique", 0.8), ("yoga", 0.4)}*,

- la **sémantique associée au contenu** : la sémantique de contenu permet de donner plus de détails concernant un *ElementDeContenu*. Ainsi, un *ElementDeContenu* est décrit par un type de valeur qui peut être un type standard (chaîne, entier…), un type dérivé tel que des unités par exemple (« cm », « pouce »…), ou encore un type qualitatif (exemple « grand »). Dans ce dernier cas, la valeur qualitative « grand » pourra être explicitée par une expression (telle que « >=30 ») reposant éventuellement sur des opérateurs

logiques (ET, OU, NON). Les *liensTypesDeValeur* permettent d'établir des relations entre les types utilisés (exemple : pouce → entier).

On peut souligner le rôle important que joue le niveau sémantique. En effet, il décrit précisément les éléments composant un profil. Il permettra de réaliser un échange entre les différentes applications même si les modèles utilisés sont hétérogènes (en termes de structure notamment).

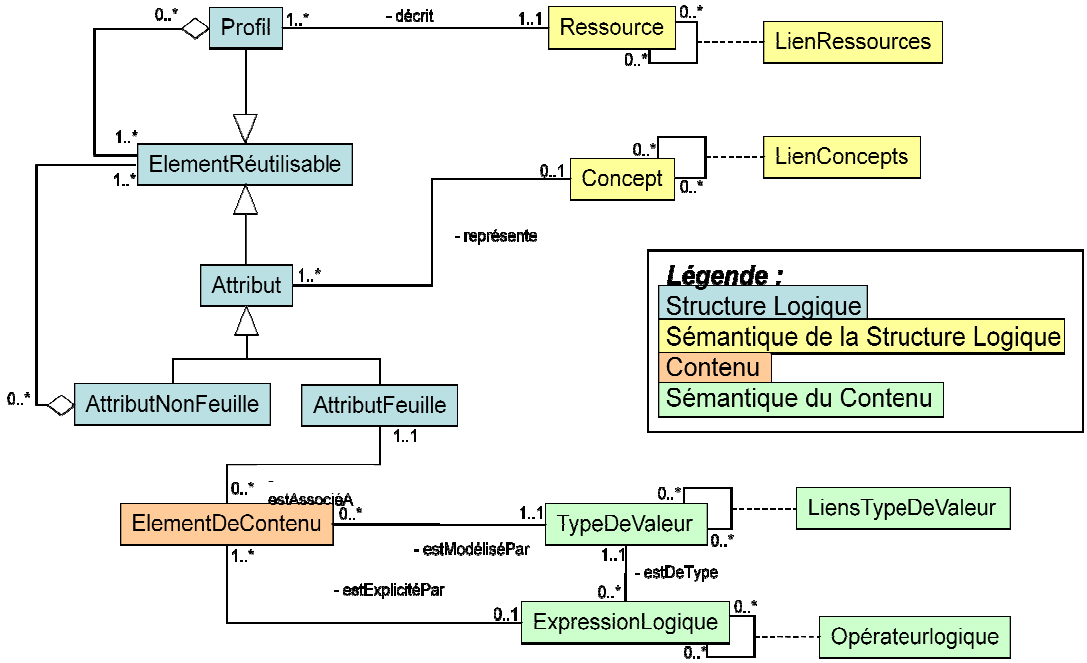


Figure 45 - Proposition de modèle de profils de ressources flexible

Cette structure sémantique fournit les éléments du modèle pivot. Ainsi ces différents éléments sémantiques dépendent du domaine d'application et peuvent être choisis librement. Ce niveau sémantique partagé entre les applications doit résulter d'un consensus entre les différentes applications. L'hétérogénéité et la variété des profils identifiés dans la littérature nous fait penser qu'un consensus sur le niveau sémantique est plus facile à obtenir qu'un consensus sur le contenu d'un profil à proprement parler.

En complément de cette structure sémantique partagée, une bibliothèque de fonctions, permettant par exemple la conversion entre les différents types, est fournie et peut être complétée pour chaque domaine d'application. Elle traduit un ensemble de connaissances de base relatives au domaine d'application.

E.3.2 Identification des appariements possibles

Grâce à ce modèle, les différentes applications peuvent modéliser leurs profils selon leurs besoins. Le fait qu'elles rattachent les éléments composant les profils à la structure sémantique partagée permet au système de comprendre ce que contient chacun des éléments du profil et ce pour tous les systèmes (Figure 46).

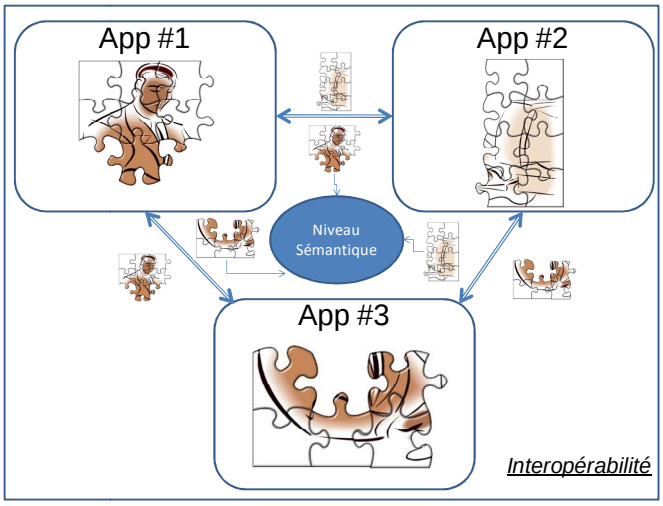


Figure 46 - Approche proposée basée sur un niveau sémantique partagé

Supposons alors le cas suivant : une application souhaite obtenir un complément de profils auprès des autres applications. Il lui suffit d'envoyer une requête en précisant le profil (*p*) qu'elle souhaite enrichir. Chacun des profils étant associé au niveau sémantique, un système implantant cette architecture est capable d'identifier, par une phase de raisonnement sur la base des différentes relations, les attributs compatibles (appariements) entre le profil *p* et chaque profil détenu au niveau de chaque application. Cette phase d'identification des appariements possibles est un passage obligé pour l'échange entre applications.

Pour obtenir les appariements possibles entre deux profils, nous avons défini six étapes (Chevalier, et al., 2006). Ces étapes sont présentées dans la Figure 47. Les deux premières étapes consistent à identifier et à vérifier la compatibilité (sémantique) entre les éléments de la structure logique. Sur cette base, les quatre autres étapes poursuivent l'exploration en vérifiant la compatibilité (à la conversion près) des valeurs de ces éléments.

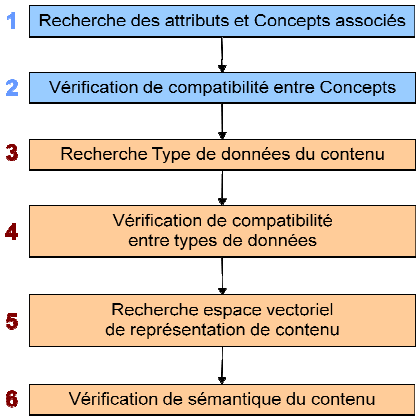


Figure 47 - Les six étapes permettant d'établir les appariements possibles entre deux profils

Finalement, sur la base des appariements identifiés entre des attributs d'un profil *q* et les attributs de *p*, l'application initiatrice de la demande peut recevoir les valeurs des attributs de *q* appariés. Cela ne nécessite pas de traitement supplémentaire puisque les valeurs des attributs de *q* ont été, si besoin, convertis au cours des six étapes dans un format identique à celui des attributs de *p* avec lesquels ils sont appariés. Pour compléter ces possibilités, nous avons également proposé, toujours sur la base des appariements identifiés

de calculer automatiquement une similarité entre deux profils (Chevalier, et al., 2006b). Pour cela, l'application initiatrice peut fixer le poids de chaque attribut du profil dans le calcul de la similarité sous la forme d'un rang par exemple.

Afin de valider l'approche, une expérimentation a été menée sur différentes collections de documents utilisées dans la campagne d'évaluation CLEF 2001 (ATS 94, LeMonde 94, LaTimes94). Pour chacun des documents, un profil a été créé à partir de la structure logique qu'il possède (les attributs ne possèdent pas le même nom et ne sont pas organisés de la même manière dans les différentes collections) et un ensemble de profils usagers ont été créés pour l'occasion. L'objectif de cette expérimentation était de vérifier si la méthode proposée permettait d'identifier automatiquement des appariements entre des profils hétérogènes. Nous avons donc comparé notre technique de détection d'appariement avec deux autres techniques utilisant soit uniquement les couples « attribut=valeur », soit uniquement la structure logique pour détecter les appariements. Les résultats obtenus montrent que la démarche basée sur la sémantique et les six étapes proposées permet d'obtenir plus de quatre fois plus d'appariements que les autres méthodes, soulignant ainsi l'intérêt de l'architecture proposée dans un contexte hétérogène.

Ces résultats nous permettent de mettre en exergue les points forts de l'architecture proposée qui est :

- **flexible** : elle permet de modéliser tous types de ressources,
- **interopérable** : elle permet des échanges d'informations sur la base d'appariements identifiés grâce au niveau sémantique partagé entre profils hétérogènes (tant en termes de contenu que de structure) notamment développés au sein plusieurs applications,
- **générique** : elle peut être instanciée dans différents contextes applicatifs tels que la RI.

E.3.3 Prototype

Afin de valider l'architecture de profils proposées, nous l'avons instanciée dans le contexte de la RI. Cette implantation a été réalisée en Java avec l'aide de l'API Jena[24] pour effectuer le raisonnement et identifier les appariements possibles entre deux profils. Les relations entre les différents éléments du modèle sont stockées sous la forme de triplets RDF (Resource Description Framework) développé par le W3C[25] et interrogés en SparQL.

Une interface de visualisation (Figure 48) a également été implantée permettant à un usager de visualiser graphiquement les appariements possibles. Dans cette figure, les flèches qui se rejoignent dans la zone centrale (correspondant au niveau sémantique partagé)

[24] http://jena.sourceforge.net/

[25] http://www.w3.org/2001/sw/wiki/Main_Page

traduisent un appariement possible entre un attribut du profil de gauche et un attribut du profil de droite.

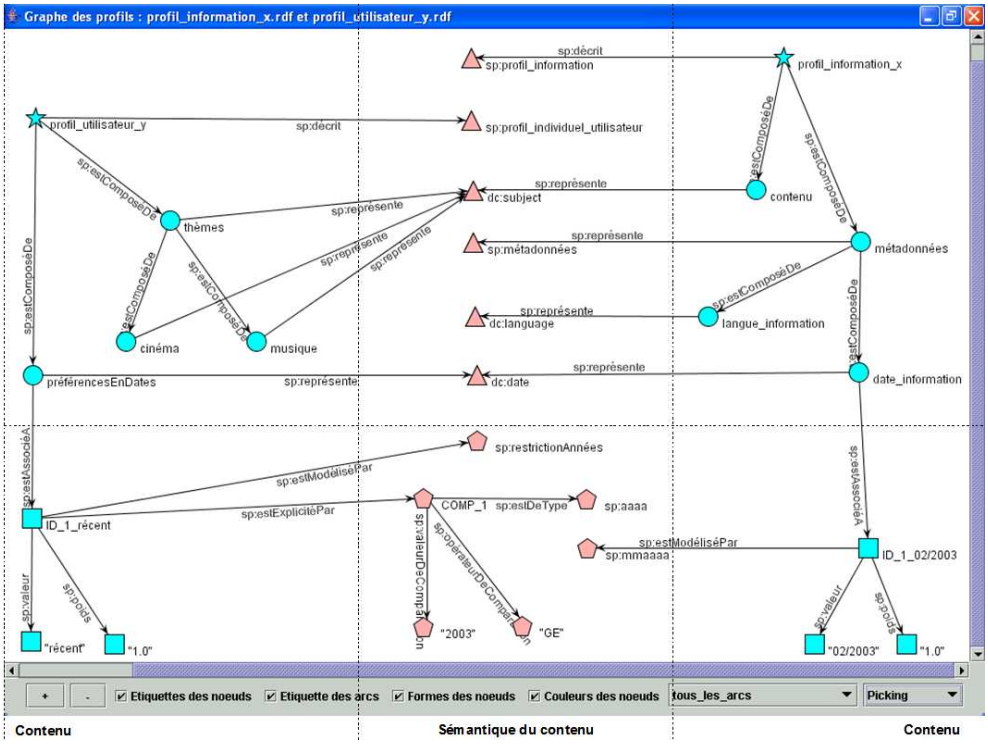


Figure 48 - Interface de visualisation des appariements possibles entre profils

Le modèle proposé permet de modéliser toutes ressources et d'échanger tout ou partie des profils entre diverses applications. Ces profils peuvent être utilisés dans le modèle de contexte pour la RI que nous proposons dans la section suivante.

E.4 Modèle de Contexte pour la RI

Les profils ainsi que les traitements associés font généralement partie intégrante de l'application, c'est-à-dire qu'il est difficile de distinguer les traitements liés aux profils des traitements autres que réalise l'application car ils sont entrelacés dans le code. Pourtant (Kobsa, 2007) souligne l'intérêt de modéliser individuellement un composant permettant la gestion des profils dans l'application.

Ces travaux (Finin, 1989), (Brajnik, et al., 1994), (Vergara, 1994), (Kay, 1995) se situent dans la mouvance des « User Modeling Shell » (UMS) et reposent principalement sur :

- une dimension de modélisation (règles, attributs-valeurs…),
- un mécanisme de raisonnement propre au modèle usager,
- un ensemble de services permettant à l'application de fournir ou obtenir des informations relatives à l'usager.

Nous estimons cette architecture pertinente pour nos travaux car elle favorise :

- la **réutilisation** : un composant pourra être répliqué dans différentes applications,

- le **découpage fonctionnel** : les informations ainsi que les traitements ne sont pas mélangés avec les traitements spécifiques de l'application. La dimension modélisation est donc clairement identifiée dans l'application,

- **l'évolution** et la **maintenance** du système : un composant pourra aisément être remplacé par un autre offrant les mêmes services,

Selon la même orientation que les travaux relatifs aux UMS, nous proposons (Chaker, et al., 2010), (Chaker, et al., 2010b), (Chaker, et al., 2010c) une architecture reposant sur un composant « gestionnaire de contexte ». L'objectif de ce composant est de fournir à l'application des informations sur les éléments constituant le contexte. Ainsi, ce composant permet notamment à l'application d'obtenir des informations non plus relatives à l'usager mais relatives à l'usager contextualisé (impacté par son propre contexte → notion de récursivité du contexte).

E.4.1 Architecture

Pour pouvoir atteindre cet objectif, le composant complexe de gestion du contexte est constitué :

- d'un composant de **modélisation**. L'application n'accède plus directement aux profils car ceux-ci ne sont pas nécessairement corrects car modifiés par le contexte. Ainsi, le point d'entrée aux profils se fait nécessairement par le composant « gestionnaire du contexte ». Tous les services initialement fournis par le composant de modélisation seront reproduis au niveau du gestionnaire de contexte.

- d'un composant de **mise en situation** (MES) des différents objets constituant le contexte (Figure 49).

Le contexte présenté dans la Figure 49 est celui de la RI qui comprend les profils hiérarchiques concernant :

- l'**usager**. Il est décrit en fonction des principales dimensions soulignées dans les chapitres précédents,
- la **tâche métier**. Ce profil permet de décrire la tâche réalisée en précisant par exemple ses sous-tâches etc…,
- l'**environnement** au sens large (température, matériel à disposition…) qui permet de compléter la caractérisation du contexte.

Le composant de Mise En Situation (MES) repose sur une base de connaissances constituée d'un ensemble de règles fourni par le concepteur de l'application et complété automatiquement par la recherche de règles d'association (Agrawal, et al., 1994). Son objectif est de construire la situation actuelle de la recherche.

Une **situation**, dans notre approche, correspond à une photographie des différents éléments constituant le contexte à un instant donné. Pour tenir compte de l'aspect récursif de la notion de contexte, ces éléments sont eux-mêmes contextualisés avant d'être intégrés à la situation.

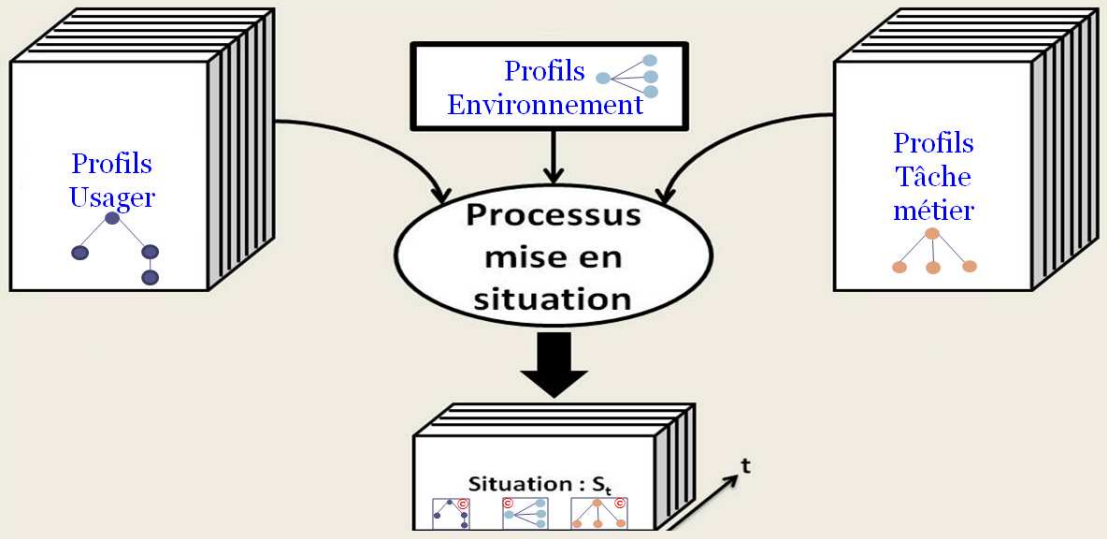


Figure 49 - Le processus de mise en situation (génération des situations)

Pour obtenir une situation, l'approche consiste à appliquer un traitement itératif à partir de l'ensemble des profils initiaux composant le contexte notés $P=\{P_{U,0}, P_{E,0}, P_{T,0}\}$. L'indice du profil correspond au type de profil suivi du numéro d'itération : U correspond à l'usager, E à l'environnement, T à la tâche.

Ce traitement repose sur trois étapes :

1) Création d'une copie de l'ensemble P dans P' dans lequel le numéro d'itération est incrémenté,
2) Pour chaque profil de P', application de toutes les règles disponibles (c'est-à-dire ayant une conclusion concernant le profil traité) en veillant à ce que chaque profil soit modifié en fonction des autres éléments du contexte de la même itération. Pour cela, chaque règle appliquée à un profil recherche la prémisse dans P alors que la conclusion est appliquée sur P'. Ainsi, pour une même itération tous les profils obtenus (situés dans P') sont modifiés par rapport aux mêmes éléments (issus de P).
3) Remplacement du contenu de P par celui de P'.

Ce processus est répété tant qu'il y a des modifications sur au moins l'un des profils ou si un cycle est détecté. En effet, la stabilité des profils est essentielle pour que la situation soit la plus fidèle possible à la réalité. Cependant, dans certains cas, certaines règles peuvent impliquer des cycles dans les traitements (en annulant l'effet d'une règle antérieure par exemple) ne permettant pas d'obtenir une stabilité de la situation. Dans ce cas, nous sommes contraints d'interrompre le traitement et la situation est stockée en l'état.

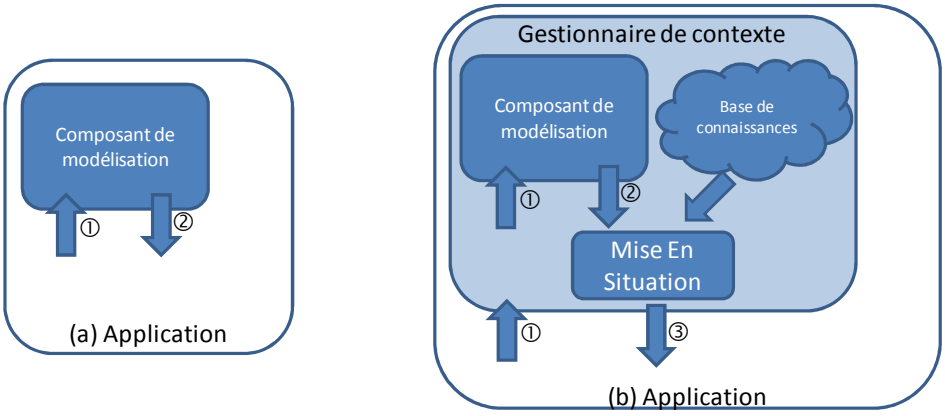
A titre d'exemple, une règle issue de la base de connaissances peut être :

Tâche.id= « soutenance » ∧ Environnement.public= « nombreux » → Usager.EtatEmotionnel = « Stressé »

Cette règle, fruit d'une imagination débordante, souligne bien l'incidence que peut avoir le contexte sur l'objet auquel il se confère. Ici, nous prenons l'exemple du contexte de l'usager qui est composé de la tâche et de l'environnement. Dans la situation, après application de la règle précédente, le profil usager possédera l'attribut *EtatEmotionnel* à la valeur « *Stressé* ».

Afin de rendre la situation la plus fidèle possible à la réalité, nous avons associé à chaque règle une priorité permettant d'optimiser le traitement de contextualisation. Dans le cas d'une règle extraite automatiquement, le poids peut être fonction de la confiance et du support de celle-ci. Pour obtenir des règles pertinentes, l'extraction des règles d'association peut être réalisée sur la base de l'historique des différentes situations passées permettant ainsi d'identifier des régularités entre les valeurs des différents profils.

Pour conclure, nous proposons dans la Figure 50 une vue synthétique de la proposition qui explique le passage d'une application adaptative (a) basée uniquement sur un composant de modélisation à une application adaptative sensible au contexte (b).



①Informations à court terme concernant l'interaction, l'environnement…
②Informations à long terme contenues dans le(s) profil(s)
③Informations correspondant à la situation actuelle (profils contextualisés)

Figure 50 - Passage d'une application adaptative (a) à une application adaptative sensible au contexte (b)

En conclusion, l'approche que nous proposons dans le cadre des travaux autour du contexte en RI est pertinente et originale sur différents plans :

- elle ne vise pas à décrire ce que doit contenir un contexte comme la plupart des travaux existants. Elle laisse l'application définir les éléments qu'elle souhaite utiliser pour caractériser le contexte. Elle propose cependant une **méthode** de mise en situation au travers de laquelle le contexte peut être inféré et peut être mis à disposition de l'application,
- la notion de **situation**, qui est le cœur de l'approche, permet à une application d'obtenir une photographie des différents éléments du contexte. Une telle situation est utile pour toute application adaptative sensible au contexte,
- elle est **générique** dans le sens où elle peut être déployée dans différents contextes applicatifs, la méthode de construction de la situation restant identique.

E.4.2 Prototype

Le modèle de contexte métier a été implanté. Cependant, il ne dispose pas encore d'interface graphique (frontal). C'est pour cette raison que nous ne pouvons fournir de copies d'écran. Toutefois, nous pouvons attester du fonctionnement général du gestionnaire de contexte.

Les profils étant généralement construits sur la base d'une hiérarchie, nous avons retenu le format XML pour le stockage des profils et pour généraliser l'utilisation du gestionnaire. Par ailleurs, le contenu du contexte n'est pas figé et peut être adapté voire totalement revu en fonction des besoins. XML joue dans ce cadre tout son rôle.

Dans cette première implantation, nous avons fait le choix de modéliser les différents profils de la manière suivante :

- le profil usager repose sur un profil hiérarchique contenant les dimensions classiques soulignées dans les chapitres précédents,
- le profil de tâche retenu est CTT (Paterno, 2000). En effet, suite à l'étude des différentes approches traitant de la modélisation des tâches et ce dans différents domaines, nous avons retenu l'approche CTT principalement car elle est extensible et associée à des logiciels de modélisation performants. Un autre aspect intéressant est que CTT permet de stocker les modèles de tâches au format XML,
- le profil d'environnement retenu est un ensemble d'attributs-valeurs.

Le gestionnaire permet de renvoyer la situation de la recherche d'information, c'est-à-dire les profils contextualisés de l'usager, de l'environnement et de la tâche. Sur la base de ces profils contextualisés, l'application peut alors réaliser les adaptations nécessaires comme décrit dans la section B.2.

E.5 Bilan et résultats

Ce chapitre a abordé l'épineuse question de la modélisation des ressources pour la personnalisation en RI. Il traite ainsi de deux problématiques :

- la modélisation dans un cadre multi-applications et plus particulièrement le problème de la « complétude » des profils,
- la modélisation et la mise à disposition du contexte aux systèmes.

E.5.1 Modélisation de profils dans un contexte multi-applications

Nos travaux ont consisté à définir une architecture qui permet aux applications de s'échanger des informations liées aux profils qu'elles possèdent. Cette architecture, est originale car elle :

- respecte la liberté des concepteurs des applications dans le choix du modèle utilisé. Ce point est d'autant plus important que les informations sur l'usager sont une réelle mine d'or pour les applications qui les possèdent,
- permet à chaque application de choisir les informations qu'elle souhaite échanger.

Cette plateforme repose sur deux éléments clés :

- un modèle flexible de profils hiérarchiques qui intègre nativement le modèle vectoriel qui est sans doute l'un des modèles les plus utilisés aujourd'hui en RI,
- une couche sémantique à laquelle les applications rattachent leurs profils. Cette couche sémantique permet au système, sur la base d'un traitement composé de six étapes, d'identifier des points d'appariement entre deux profils n'ayant pas la même structure.

Une plateforme a été implantée permettant d'évaluer ces propositions. Les résultats obtenus soulignent que l'approche proposée permet bien d'identifier des appariements entre deux profils hétérogènes.

E.5.2 Modèle de contexte pour la RI

Les profils usager ne sont que la première étape dans l'adaptation. Une tendance plus globale se focalise désormais sur la notion de contexte. Au regard d'une revue de littérature et de l'inexistence réelle de définition de cette notion, nous avons identifié certaines caractéristiques et proposé notre définition du contexte : « *Le contexte d'un objet correspond à un ensemble d'éléments pouvant avoir un impact sur cet objet* ».

Par ailleurs, la plupart des travaux tentent d'identifier ce que doit contenir un contexte pour ensuite l'exploiter. Nous ne nous sommes pas inscrits dans cette démarche car de notre point de vue, son contenu est « contextuel », c'est-à-dire qu'il dépend de l'application et du contexte dont on parle (le contexte de l'usager est différent de celui de la RI). Ceci explique l'impossibilité d'obtenir une vision unique du contexte et ce même pour le contexte de la RI.

Notre démarche a été motivée par le fait qu'un même usager par exemple dans deux contextes différents pouvait avoir des caractéristiques différentes (tant en termes de structure que de valeur). Ainsi, nous avons abordé le problème de la mise à disposition du contexte à l'application souhaitant adapter la RI par exemple. En effet, les éléments du contexte définis par l'application ne sont pas statiques car ils possèdent un contexte qui peut avoir un impact sur eux-mêmes. Fort de cette idée, nous avons proposé un modèle de gestion de contexte qui permet de créer et de mettre à disposition du système la situation dans laquelle il se trouve : elle correspond à une photographie réalisée à un instant donné du contexte du système.

Pour obtenir cette photographie, une base de connaissances est utilisée (règles fournies par le concepteur ou extraites automatiquement de l'historique des situations). Les règles extraites automatiquement sont issues de l'historique des situations.

En conclusion, nous avons identifié les principales leçons tirées de ces travaux :

- il est important que chaque application puisse librement modéliser l'usager ou les autres objets qu'elle manipule de la façon qui lui convient. En effet, surtout dans le contexte du Web, une vision unique et surtout « imposée » risque de limiter les chances d'acceptabilité des solutions : l'interopérabilité trouve naturellement sa place dans pareilles situations,
- la notion de contexte n'est pas évidente à appréhender puisque sa définition est spécifique à chaque application. Dans ce cadre, nous nous sommes intéressés à l'impact des éléments qui composent un contexte et ce quels que soient sa représentation et son contenu. Pour cela un travail mené conjointement notamment avec des chercheurs en SHS (psychologie, sciences cognitives, ergonomie) est nécessaire pour comprendre l'impact du contexte sur l'usager (état motivationnel, émotionnel, de frustration…).

E.6 Valorisation et support à la recherche[26]

Ces travaux servent de support à deux thèses (dont une soutenue) et un stage de Master 2. Ils ont donné lieu à 11 publications.

Encadrement	Publications				Projets associés
	RI	RN	CI/WI	CN	
2 thèses (P. Tchienehom, H. Chaker) et 1 M2 (A. Moussawe)		2	5	4	*AmIE, APMD*

Les propositions relatives au modèle de ressources interopérable sur la base du niveau sémantique partagé ont été développées dans la thèse de Pascaline Tchienehom soutenue en 2006. Ces travaux ont été notamment réalisés dans le cadre du projet APMD.

Dans le même objectif, le stage de Master 2 de Ahmad Moussawe visait à étudier la possibilité et à implanter un système d'échange de profils usagers entre deux systèmes de recherche d'information. Ce sujet était dans la continuité des travaux de thèse de Pascaline Tchienehom ainsi que de (Chevalier, et al., 2008b).

La thèse de Hamdi Chaker étend les travaux précédents en les orientant sur le « contexte ». Elle s'inscrit dans les motivations du projet AmIE dans lequel le « contexte » tient une place importante pour les entités ambiantes. L'objectif de cette thèse est de définir une plateforme de gestion de contextes génériques.

E.7 Perspectives sur les contributions

Nous avons identifié un ensemble de perspectives directement liées à chaque proposition.

E.7.1 Modèle de ressources interopérable

Le modèle de ressources interopérables a été défini sur la base notamment du modèle vectoriel qui est un modèle communément utilisé en recherche d'information. Une première perspective serait **l'extension** de ce modèle pour qu'il accepte d'autres types de modèles (autres que le modèle vectoriel).

Par ailleurs, les problèmes d'inconsistances et d'incohérences entre les valeurs échangées doivent être abordés. Une perspective concerne la **traçabilité des informations** liées à un usager et qui seront échangées. Ce point doit être particulièrement étudié pour que chaque application puisse comprendre d'où vient réellement l'information et quelles sont ses caractéristiques (ex : date de dernière mise-à-jour). Sans cela, les informations échangées ne pourront pas être intégrées par les systèmes et seraient donc inutiles.

Enfin, une dernière perspective concerne **l'identification des usagers** au sein des différentes applications. Les travaux proposés ne peuvent fonctionner que si les applications sont capables d'identifier de manière unique, avec un même identifiant, chaque usager. Une étude approfondie des systèmes d'identification partagée tels que *OpenId*[27] est donc requise

[26] Plus de détails concernant les thèses et Master 2 encadrés sont fournis en section F.1 (p. 103).

[27] http://openid.net/

comme point de départ de nouvelles propositions dans ce cadre. L'extension de tels systèmes favoriseront donc les échanges entre les différentes applications.

E.7.2 Modèle de contexte

Les perspectives relatives au modèle de contexte sont triples. La première concerne l'application récursive du contexte sur les objets. Une étude plus approfondie des problèmes liés aux **cycles** dans l'exécution des règles doit être menée afin de proposer une situation la plus fidèle possible à la réalité.

La seconde concerne **l'extraction automatique des règles**. En effet, les variables d'un profil peuvent ne pas être mono-valuées (attribut=valeur). Dans le cas de la recherche d'information par exemple, les attributs peuvent être multi-valués (attribut={$valeur_1$, $valeur_2$…$valeur_n$}) du fait notamment de l'emploi du modèle vectoriel. Cette perspective vise donc à insérer une similarité de contenu dans l'algorithme d'extraction des règles pour indiquer que deux variables sont suffisamment similaires pour être considérées comme identiques.

Enfin, la troisième perspective concerne **l'extension de la base de connaissances** utilisée. Ainsi, en plus de la situation, il serait pertinent d'intégrer des informations liées à l'activité de l'usager au travers par exemple d'un log d'interaction. Cette nouvelle connaissance permettrait au système d'indiquer de façon anticipée les actions régulières que réalise l'usager dans le contexte. Ces informations permettraient au SRI d'adapter encore plus finement le processus même de recherche d'information.

E.8 Perspective sur les modèles en recherche d'information

La principale perspective que j'envisage concerne les problèmes d'accès aux informations concernant un usager (ou plus largement un objet) détenues par les systèmes. En effet, les systèmes doivent respecter la **confidentialité et la sécurité** des données. Certains travaux tels que (Viviani, et al., 2010) peuvent servir de point de départ à des propositions nouvelles favorisant interopérabilité et sécurité. Par exemple une **implication de l'usager** plus importante dans l'exploitation des données (afin qu'il puisse valider ou non l'utilisation de ses données) pourrait être une réponse à ce problème. L'objectif de ces travaux vise à instaurer une réelle relation de *confiance* entre le système et l'usager.

F. VALORISATION DES TRAVAUX

Les sections suivantes présentent la structuration et la diffusion de la recherche que j'ai assurée ou encadrée depuis 2002.

F.1 Formation à la recherche

Depuis 2002, j'ai organisé mes travaux pour favoriser le support à la recherche pour des étudiants doctorants ou en Master 2. Ainsi, j'ai assuré ou j'assure le co-encadrement scientifique de :

- 5 thèses dont 2 soutenues (2006 et 2008),
- 3 stages de M2.

	Nombre	%Encadrement
Encadrement d'étudiants en Master 2	***3***	
Encadrement d'étudiants en Thèse	***5***	
Ayant soutenu	2	50%, 60%
En cours	3	50%, 50%, 50%

Figure 51 - Récapitulatif de l'encadrement que j'ai effectué depuis 2002

F.1.1 Thèses informatiques encadrées

DAMIEN DUDOGNON, « **Analyse des mouvements d'information et fouille de textes - Stream analysis and text mining** »,
Thèse de Doctorat de l'Université Paul Sabatier - Toulouse 3.
Financement CIFRE (contrat 1274/2009).
Janvier 2010 - soutenance prévue en 2013.
Encadrement à 50% - Directeur de Recherche : J. Mothe
Cette thèse a pour objectif de développer des approches basées sur l'activité des usagers autour des informations disponibles et notamment autour des blogs. Les travaux menés dans cette thèse sont réalisés dans le cadre du projet OverBlog.

HAMDI CHAKER, « **Modélisation et exploitation de profils : accès sémantique a des ressources** »,
Thèse de Doctorat de l'Université Paul Sabatier - Toulouse 3.
Financement MENRT (thème prioritaire).
Septembre 2008 - 2011.
Encadrement à 50% - Directeurs de Recherche : C. Soulé-Dupuy, A. Tricot
Cette thèse a pour objectif de développer un modèle de contexte pour l'amélioration des systèmes de recherche d'information. Cette thèse peut être considérée comme une extension de la thèse de Pascaline Tchienehom. Ce travail est développé dans le cadre du projet AmIE.

CHRISTIAN BELBEZE, « **Communautés dynamiques et portails collaboratifs** »,
Thèse de Doctorat de l'Université des Sciences Sociales - Toulouse 1.
Formation continue (Salarié).
Septembre 2007 - soutenance prévue en 2011.
Encadrement à 50% - Directeur de Recherche : C. Soulé-Dupuy
Cette thèse a pour principal objectif d'améliorer l'épineux problème de la formulation des besoins dans les systèmes de recherche d'information identifiés dans le projet GenIE. Après une étude des possibilités de personnalisation des systèmes de recherche d'information (moteur à base de composants), cette thèse repose sur une dimension collective ainsi que sur les approches de personnalisation pour désambigüiser dynamiquement les requêtes des usagers.

GUILLAUME CABANAC, « **Fédération et amélioration des activités documentaires par la pratique d'annotation collective** »,
Thèse de Doctorat de l'Université Paul Sabatier - Toulouse III.
Financement MENRT.
Septembre 2005 - Décembre 2008.
Encadrement à 60% - Directeur de Recherche : C. Chrisment, co-encadrement avec C. Julien
Devenir : depuis 2009, Maître de conférences à l'IUT A Rangueil, Département Informatique
Cette thèse s'intéressait particulièrement aux processus pour enrichir les activités liées aux documents électroniques. Elle abordait également la visualisation multi-facettes des résultats de recherche. Cette thèse peut être considérée comme une extension de ma thèse soutenue en 2002.

PASCALINE-LAURE TCHIENEHOM, « **Modélisation et exploitation de profils : accès sémantique a des ressources** »,
Thèse de Doctorat de l'Université des Sciences Sociales - Toulouse 1.
Financement MENRT.
Septembre 2002 - Novembre 2006.
Encadrement à 50% - Directeur de Recherche : C. Soulé-Dupuy
Devenir : depuis 2006, SSII région Parisienne
Cette thèse avait pour objectif de développer un modèle de ressources interopérables. Cette thèse était en phase avec l'analyse du problème complexe de la modélisation usager en recherche d'information menée notamment au travers du projet APMD.

F.1.2 Stages de Master 2 Recherche encadrés

A. AL-MOUSAWE, « **Interopérabilité de modèles usagers dans un contexte partagé** », 2009-2010.
Directeur de Recherche : C. Soulé-Dupuy.
Thème : proposition et implantation d'un modèle usager partagé entre différents systèmes de recherche sur le Web. Ce projet est directement lié à la problématique de modélisation usager que nous développons dans nos travaux (thèse de Pascaline Tchienehom).

R. AMARA, « **Web services et Accès à l'information** », 2005-2006.
Directeur de Recherche : C. Soulé-Dupuy.
Thème : étude et modélisation d'un système basé sur les Web services et permettant d'améliorer l'accès à l'information. L'orientation de ce projet a permis de mieux appréhender les problématiques développées dans la thèse de Christian Belbèze.

G. CABANAC, « **Annotations de ressources électroniques sur le Web : formes et usages** », 2004-2005.
Directeur de Recherche : C. Chrisment, co-encadrement avec C. Julien.
Thème : étude et réalisation d'un prototype permettant d'annoter et partager des annotations sur des documents électroniques du Web. Ce projet a été le point de départ de la thèse du même auteur débutée en 2005.

F.2 Publications relatives à nos travaux (depuis 2002)

Les travaux présentés dans ce mémoire ont été validés par des publications scientifiques dont nous donnons un aperçu synthétique dans le Tableau 7.

Nature des manifestations	Nombre
Revues internationales	1
Revues nationales	6
Edition d'ouvrages collectifs internationaux	1
Edition de numéro spécial de revues internationales (appel à communication en cours)	1
Contribution à des ouvrages collectifs internationaux	2
Contribution à des ouvrages collectifs nationaux	2
Conférences & Workshop internationaux avec comité de sélection et actes publiés	16
Conférences & Workshop nationaux et francophones avec comité de sélection et actes publiés	15
Autres publications	6

Tableau 7 - Tableau des publications relatives aux travaux menés depuis 2003

Une ventilation des principales publications selon les différents axes de recherche est fournie dans le tableau synthétique (Tableau 8). La liste complète des publications est fournie dans le chapitre H.

F.3 Projets et Collaborations Scientifiques

L'orientation des recherches dans lesquelles je m'inscris depuis 2002, du fait qu'elles soient centrées sur l'usager et la RI, touche de multiples domaines et champs d'applications. Ainsi pour obtenir une vision plus globale de l'usager et des problèmes auxquels il est confronté, nous collaborons avec des partenaires industriels et chercheurs dans le cadre de projets ou pour assurer la diffusion scientifique de nos propositions ou tout simplement pour échanger afin de partager notre expérience et expertise dans nos domaines d'intérêt.

Les échanges enrichissants que nous avons eus jusqu'à présent ont permis la maturation des travaux dont nous exposons les résultats dans ce mémoire. Ainsi nous nous sommes enrichis de collaborations (régionales, nationales et internationales) avec des partenaires dans les domaines de l'**Informatique**, des **Sciences Humaines et Sociales** (plus précisément en sciences cognitives), des **Mathématiques** (plus précisément en théorie des graphes), de la **Gestion** et de la **Biologie**.

Ces partenariats ont donné lieu à différents projets et collaborations industrielles :

- 2010-2013 : **OverBlog**

Cadre :

Collaboration industrielle - CIFRE (contrat ANRT 1274/2009)

Partenaires :

Société JFG Networks, IRIT

Objectifs :

Nous avons établi au travers de la bourse CIFRE de Damien Dudognon une collaboration industrielle avec la société JFG Network/Overblog. L'objectif de ce projet est de caractériser les usagers, les informations et les relations qu'ils entretiennent entre eux (usagers/usagers, Information/Information, Usagers/Information). Une seconde étape est d'identifier et caractériser les mouvements (émergence, « buzz ») autour de l'information disponible dans

des blogs. Cette caractérisation de la dynamique de l'information revêt un intérêt tout particulier pour la recherche d'information car elle nous permettra de compléter la notion de pertinence système.

Implication personnelle (15 % du temps consacré à la recherche) :

Participant, encadrant de la thèse de Damien Dudognon. J'apporte à ce projet mes compétences dans les domaines de la **recherche d'information** et de la **personnalisation** notamment pour proposer des approches permettant d'améliorer l'accès à l'information par la modélisation de l'activité globale autour des blogs. Ce projet apporte un **cadre d'expérimentation** intéressant pour valider nos travaux au travers d'une plateforme de plus de 1 300 000 blogs associés à une grande quantité d'informations telles que les logs d'accès (2 M de visiteurs uniques par jour), les logs de requêtes…

- 2009-2010 : **Ambiant Intelligent Entities (AmIE)**

Cadre :

BQR UPS

Partenaires :

9 équipes de l'IRIT

Objectifs :

L'objectif de ce projet est triple :

- Fédérer les recherches autour des systèmes ambiants,
- Définir et modéliser le socle générique d'une architecture interopérable, ouverte et extensible pour l'ambiant,
- Développer ou adapter des briques logicielles complémentaires, mutualisables et réutilisables, indispensables pour la réalisation d'applications émergentes novatrices pour utilisateurs nomades.

Implication personnelle (20 % du temps consacré à la recherche) :

Participant. Dans le cadre de ce projet, notre implication concerne principalement la modélisation des entités ambiantes (usager, entité ainsi que de leur contexte). La dimension interopérabilité entre ces modèles est également présente dans ce projet. J'ai donc apporté ma contribution au niveau des **modèles interopérables de ressources** (thèse de Pascaline Tchienehom) et plus particulièrement de la **définition du contexte** (thèse de Hamdi Chaker). Mes compétences en système d'informations me permettent également d'avoir une vision plus large des contraintes liées à l'usage des informations que l'on peut développer dans le domaine de l'ambiant (systèmes de recommandation, classifieurs…). Ce projet permet également d'appliquer mes travaux dans un champ d'application différent permettant ainsi d'en identifier les limites et les éventuelles évolutions.

- 2004-2007 : **Accès Personnalisé à des Masses de Données (APMD)**

Cadre :

ACI Masse de Données (Projet MD-33)

Partenaires :

PRISM, LIG (CLIPS-IMAG), IRISA, IRIT, LINA, LIRIS

Objectifs :

Le projet APMD a pour objectif de mener une réflexion globale sur la personnalisation de l'information dans un environnement à grande échelle. Plus précisément, l'objectif est de proposer des modèles formels capables de capturer les besoins des utilisateurs et de les représenter dans des profils qui sont utilisés par des algorithmes robustes pour un accès et une présentation adaptative de l'information. Le projet est bâti autour de trois dimensions : la notion de profil et son mode de représentation et d'élaboration, la qualité de l'information et ses méthodes de mesure, le modèle d'exécution adaptatif des requêtes basé sur les profils.

Implication personnelle (20 % du temps consacré à la recherche) :

Participant. Notre participation se situe précisément au niveau de la modélisation de l'usager. Ce projet a notamment permis la **synthèse des indicateurs pour la modélisation des profils** pour la recherche d'information et m'a permis de mieux appréhender les problèmes sous-jacents à la modélisation usager. Il m'a aidé à **développer la vision du modèle interopérable** que nous avons développé. Ainsi, les publications relatives à la définition de ce modèle interopérable de ressources rentrent dans le cadre de ce projet (thèse de Pascaline Tchienehom).

- 2005-2006 : **Genomic Information Extraction (GenIE)**

Cadre :

BQR UPS

Partenaires :

IRIT, Laboratoire de Statistiques et Probabilités, INRA, LibraGen, Cayla, Millegen

Objectifs :

L'objectif de ce projet a été d'identifier et de donner des pistes de solutions pour la recherche d'information en génomique. En effet, l'hétérogénéité ainsi que la quantité d'informations fortement disséminées ne facilite pas la tâche des usagers qui doivent valider des relations que possèdent différents objets biologiques au travers des informations disponibles. Ce projet a été l'occasion d'observer les pratiques de recherche d'information dans le domaine de la génomique pour identifier les problèmes rencontrés par les usagers.

Implication personnelle (25 % du temps consacré à la recherche) :

Coordinateur du projet. Dans ce projet, j'ai pris en charge la **rédaction** de la réponse à l'appel à projet et celle du rapport final. Ce projet a été l'occasion de me **sensibiliser** aux problèmes de recherche d'information qui s'inscrit dans une activité plus complexe : l'identification de liens entre les objets biologiques provenant de différentes sources d'informations. Dans ce cadre, j'ai encadré un stagiaire qui a eu la tâche de développer un moteur de recherche « étendu » qui permet d'automatiser la recherche des relations

entre les objets d'intérêt dans un ensemble de sources de connaissances génomiques identifiées. Les résultats de ce projet m'ont permis d'identifier les problèmes de formulation des besoins et d'y pallier en proposant un moteur de recherche « personnalisé ». Ces travaux sont développés dans le cadre de la thèse de Christian Belbèze.

F.4 Tableau synthétique de la structuration et la valorisation de nos travaux

Axes de Recherche	Encadrement*	Publications				Projets associés
		RI	RN	CI/WI	CN	
Axe #1 : fédération des activités documentaires	2 thèses (G. Cabanac, D. Dudognon) et 1 M2 (G. Cabanac)	1	2	7	5	*Overblog*
Axe #2 : Amélioration des SRI centrée Usager	1 thèse (C. Belbèze) et 1 M2 (R. Amara)		2	4	6	*GenIE*
Axe #3 : Modélisation pour la personnalisation en RI	2 thèses (P. Tchienehom, H. Chaker) et 1 M2 (A. Moussawe)		2	5	4	*AmIE, APMD*

* Chaque thèse est positionnée dans l'axe le plus développé

Légende : RI : revue internationale
RN : revue nationale
CI/WI : conférence internationale / workshop international
CN : conférence nationale

Tableau 8 - Tableau synthétique de la structuration et de la valorisation de nos travaux

F.5 Animation Scientifique autour de nos problématiques de recherche

En parallèle à mes activités de recherche, je veille à participer à l'animation scientifique autour des thématiques de l'usager et de la recherche d'information, ainsi :

- Je fais partie depuis 2009 de l'ensemble des animateurs du thème 3 (Ingénierie par et pour les modèles dans les systèmes d'information) du GDR I3. Je participe plus précisément à l'animation autour de la thématique de la prise en compte des usagers dans les systèmes d'information,
- Je participe en qualité de co-éditeur à :
  - un ouvrage collectif international autour de la modélisation usager publié en 2008,
  - un numéro spécial de revue internationale (en cours) autour de la prise en compte des usagers dans les systèmes d'information.
- Depuis 2004, je suis membre élu du bureau (chargé de communication puis trésorier) de l'ARIA (Association francophone en Recherche d'Information et Applications) qui anime la communauté nationale dans le domaine de la recherche d'information et qui organise entre autres le congrès annuel CORIA (Conférence francophone en Recherche d'Information et Applications).

G. CONCLUSION

G.1.1 Synthèse

Les travaux que j'ai pu mener et encadrer depuis 2002 ont consisté à apporter des solutions en matière d'amélioration des activités documentaires et plus particulièrement de l'activité de recherche d'information. Le leitmotiv de mes travaux est l'étude de la relation entre usagers et information disponible. J'ai articulé ces travaux autour de trois axes de recherche principaux :

- la fédération des activités documentaires visant à apporter une amélioration globale de la gestion documentaire,
- l'amélioration de l'usage des systèmes de recherche d'information,
- la personnalisation et l'adaptation en recherche d'information.

Axe 1 : Fédération des activités documentaires visant à apporter une amélioration globale de la gestion documentaire.

La recherche d'information n'est pas une fin en soi. Elle s'insère dans un processus plus large qui nécessite souvent la réalisation d'autres activités. Ainsi dans le cadre de la GED, les activités ont été présentées et nous avons souligné l'inexistence d'interaction entre ces activités.

Les travaux dans cet axe de recherche ont permis de mettre en évidence que l'interaction entre les différentes activités documentaires permet d'améliorer de façon globale la gestion documentaire. Ainsi nous avons montré une architecture de fédération des activités qui repose sur :

- un ensemble de processus adaptatifs qui permettent un enrichissement mutuel entre les différentes activités documentaires,
- une dimension collective qui permet à chacun de participer, au travers de ses activités, à l'amélioration des activités des autres usagers du système selon un principe donnant-donnant.

Par ailleurs, nous avons défini des indicateurs utiles pour l'amélioration globale de l'accès à l'information (processus Navi par exemple). Ces indicateurs sont essentiellement basés sur l'activité des usagers.

Un prototype a été développé et chacun des processus proposés évalué. Les résultats obtenus ont permis de valider les propositions.

Axe 2 : Amélioration de l'usage des systèmes de recherche d'information.

Ces travaux se sont intéressés aux problématiques de la recherche d'information basée sur le requêtage (systèmes de recherche d'information) et particulièrement sur les problèmes rencontrés par l'usager : la formulation des besoins et l'exploitation des résultats de recherche.

En résultat, nous proposons un système de médiation pour la désambigüisation de requête permettant à l'usager d'adapter, avec l'aide du système, sa requête en fonction de ses besoins. A l'autre bout du processus de recherche d'information, nous avons proposé une interface de visualisation multi-facettes permettant à l'usager d'explorer les documents selon différentes dimensions. En parallèle, nous avons proposé une plateforme de mesure de l'efficacité des interfaces de visualisation pour la recherche d'information. L'usage de cette plateforme nous permettra de répondre à la question : « quelle visualisation est plus adaptée à un contexte de recherche d'information donné ? ». La réponse à cette question permettra dans le futur de guider l'usager dans l'exploration des documents afin d'identifier les documents qui répondent à ses besoins. Enfin, nous nous sommes intéressés à un système de recherche permettant à l'usager de décrire, à l'aide de composants, une chaîne de traitements sur les résultats de recherche pour qu'il puisse pleinement les manipuler. Cette réflexion est d'autant plus importante lorsque l'usager doit réaliser une recherche complexe d'information (fusion de résultats de différents outils de recherche -texte vs image-, filtres …) ou qu'il possède des handicaps nécessitant une adaptation des résultats (ex : daltonisme).

Les propositions ont été implantées au travers de prototypes pour permettre leur évaluation. Les résultats obtenus confirment l'intérêt des propositions.

En conclusion, ces travaux nous ont permis de souligner l'importance des techniques de visualisation dans le domaine de la recherche d'information.

Axe 3 : Modèles pour la personnalisation en recherche d'information.

Cet axe de recherche concerne une dimension transversale des différentes propositions visant à répondre de façon adaptée à l'usager. En effet, la modélisation des différentes ressources (usager, périphérique…) impliquées dans un système est primordiale à des fins de personnalisation. Le fait que la recherche d'information peut être réalisée par un ensemble d'outils nous a conduit à nous intéresser à la modélisation inter-applications. Nous avons donc proposé un modèle de profils interopérables à partir duquel les applications peuvent s'échanger des informations relatives aux usagers afin de compléter leur vision de l'usager et ainsi lui apporter une réponse plus adaptée.

Dans un second temps, cette vision centrée sur l'usager a été élargie pour prendre en compte la notion de contexte. En effet, un même objet peut être caractérisé différemment dans différents contextes. Cette notion est donc importante pour la personnalisation. Nous avons proposé un modèle de contexte qui, sur la base d'un gestionnaire de contexte, livre au système de recherche une photographie fidèle des objets d'intérêt contextualisés pour l'application.

Ce gestionnaire de contexte a été implanté et peut intégrer des modèles de ressources comme celui proposé dans nos travaux.

Tout au long de nos activités de recherche nous avons donc porté une attention particulière à la validation des propositions que nous avons pu formuler. Ainsi, le **développement** de prototypes, **l'expérimentation** à des fins de validation ainsi que la **diffusion scientifique** ont été au cœur de nos propositions.

Par ailleurs, dans ces travaux, une des motivations principales a été de tenir compte de **l'hétérogénéité** à laquelle font face les outils informatiques pour la recherche d'information.

Le Tableau 9 présente les réponses originales que nous avons apportées pour chacune des hétérogénéités identifiées.

Pour ce faire, les activités que nous avons menées se situent à l'intersection de trois principaux domaines de recherche : la **recherche d'information**, la **visualisation d'information** ainsi que la **personnalisation**. Nous avons synthétisé dans la Figure 52 l'implication relative de chacun de ces domaines dans les axes de recherche que nous poursuivons.

Ces différents axes de recherche ont été abordés dans le temps de façon parallèle (Figure 53) du fait de leur complémentarité.

Enfin, ces travaux sont le fruit d'un travail structuré autour notamment de l'activité de 8 étudiants (5 thèses et 3 Master 2 recherche) pour lesquels je participe ou j'ai participé à l'encadrement. Grâce à ce travail collectif, les résultats obtenus sont positifs et nous encouragent à poursuivre nos travaux. Trois perspectives principales à nos travaux peuvent ainsi être soulignées.

G.1.2 Perspectives

Une première perspective concerne l'exploitation et la réutilisation des différentes propositions. Actuellement les différents prototypes développés sont indépendants. Afin d'organiser la recherche future, il est important de disposer d'une plateforme à base de composants permettant l'intégration de nos propositions actuelles et de nos propositions futures. Cet objectif est double :

- favoriser l'**intégration de nouveaux partenaires** dans nos activités de recherche (étudiants ou membres de l'équipe par exemple). L'approche par composants permettrait à ces nouveaux partenaires de développer leurs propres composants avant de les insérer dans le système commun,

- favoriser les **tests et les expérimentations** au plus tôt dans la recherche grâce à l'approche par composants et plus particulièrement à la possible substitution entre deux composants.

Une seconde perspective concerne la pierre angulaire de nos travaux : la recherche d'information. Dans ce cadre, une proposition consiste à développer les **systèmes de recherche d'information collaboratifs**. Plus précisément il s'agit de développer l'aspect collaboration **synchrone** durant la recherche d'information comme cela est le cas dans les dialogues interpersonnels. En effet, je pense qu'il est pertinent de s'intéresser aux meilleurs moteurs de recherche qui soient : *les usagers*. Une autre orientation possible concerne l'usage des **informations exogènes** à l'information (annotations, réseaux sociaux…) pour améliorer la recherche d'information en favorisant la convergence entre la pertinence système et la pertinence usager. Une première analyse des évolutions possibles sur la base de l'usage des annotations a été proposée dans [19]. Cependant, pour atteindre ces différents objectifs, la collaboration avec les partenaires dans le domaine des SHS notamment doit être maintenue voire développée pour nous permettre d'appréhender au mieux la démarche de l'usager lors d'une recherche d'information.

Une perspective plus ambitieuse à plus long terme concerne **l'évolution des systèmes de recherche d'information**. En effet, l'ensemble de nos travaux met en lumière le fait que les outils actuels de RI n'ont pas une portée suffisante pour permettre à l'usager de rechercher, traiter, etc. l'information retrouvée de façon à réaliser une tâche complexe. Dans ce cadre, nous proposons d'inverser la tendance en faisant **converger les systèmes de recherche d'information et les outils de Data Mining** permettant à l'usager d'explorer et de mettre en évidence des relations entre les informations retrouvées. Cependant, ces outils de Data Mining requièrent une expertise pour pouvoir être utilisés de façon adaptée et efficace que ne possèdent pas la plupart des usagers de recherche d'information surtout dans le contexte du Web. Les techniques de Data Mining doivent donc être adaptées et adaptables afin qu'elles puissent être mises à disposition de n'importe quel usager même sans expertise. Ceci nécessite de transférer les connaissances de l'expert en Data Mining au sein du système. Pour cela une base de connaissances peut être utile pour capter cette expertise et la restituer au système. Dans le même temps, il est primordial de faciliter la communication entre le système et l'usager pour le guider au mieux, si besoin, dans ses activités de recherche. Pour cela nous estimons que les approches de médiations sont plus adaptées. Ainsi, un tel système doit, pour lui permettre de progresser, reposer sur des assistants plutôt que sur des processus totalement automatisés se substituant à l'usager.

Les facteurs clés pour la réussite d'un tel challenge résident, de notre point de vue, dans :

- une **modélisation** précise de l'usager mais également des outils disponibles, des informations disponibles, de l'environnement…. Cette modélisation repose entre autres sur l'identification et la formalisation d'indicateurs (endogènes mais également exogènes à la RI) permettant de compléter la perception de la situation dans laquelle l'activité de l'usager se situe : intentionnalité, traçabilité…

- la mise en place de techniques **d'adaptation** efficaces tenant compte des différents objets mis en jeu pour permettre une amélioration de l'activité globale de l'usager,

- de **techniques de visualisation** adaptées permettant au système de communiquer les informations pertinentes à l'usager.

D'un point de vue prospectif, je compte **poursuivre nos travaux sur la personnalisation de la RI** car cette activité devrait continuer à croitre notamment dans le cadre du Web. Les problèmes d'hétérogénéité, de la masse des usagers et des informations persisteront même si nous pensons qu'un recentrage sur des informations « propres » devrait s'opérer favorisant ainsi la tâche de recherche d'information. Il est donc important de poursuivre les orientations que j'ai poursuivies depuis le début car je pense qu'**elles resteront d'actualité encore quelques années**. Le nombre de requêtes sur les moteurs de recherche ainsi que la taille du Web en constante croissance me conforte dans cette idée.

Cependant, dans ce domaine, de nouveaux challenges apparaissent et me semblent incontournables comme la recherche verticale (« vertical search ») qui permet de retrouver différents types de documents (pages web, images, vidéos…) au travers d'une même et seule requête. Pour ma part, vus les travaux que j'ai déjà entrepris, je suis plus sensible aux problèmes de modélisation des usagers dans un contexte multi-applications (surtout lorsque

celles-ci permettent de remplir la même tâche). Ainsi une évolution de mes travaux peut être vue dans les problématiques d'**identité numérique** (« digital identity ») qui semblent être un verrou qu'il faut lever pour favoriser l'émergence de nouveaux systèmes personnalisés respectant le respect de la vie privée. Un des précurseurs en la matière (Dick Hardt) confirme la nécessité de se poser la question de l'identité numérique et souligne déjà une avancée notable dans ce domaine avec par exemple le projet *OpenId* particulièrement dans le monde des blogs. Il ne s'agit plus de se limiter à l'identification unique des usagers mais d'intégrer un maximum d'informations pouvant représenter les usagers et de fournir ces informations aux systèmes de façon sécurisée. Toutefois, le cas des informations destinées à la personnalisation ne sont pas, à ma connaissance, évoquées dans l'identité numérique. Ces informations concernent pourtant bien la vision que les systèmes ont sur l'usager et il serait bon de les inclure dans l'identité numérique. Certaines d'entre elles revêtent cependant un caractère tellement **contextuel** à chaque application (préférences spécifiques à une application) que leur intégration dans l'identité numérique doit être étudiée.

Le challenge que je me fixe dans ce cadre est d'**intégrer les informations liées à la personnalisation à l'identité numérique** afin qu'elles puissent être partagées entre différentes applications permettant ainsi d'améliorer la vision que chaque système possède sur l'usager. La **recherche d'information sur le Web** reste pour moi le cadre applicatif à privilégier pour expérimenter nos propositions.

Pour atteindre cet objectif et **faire émerger cette thématique au sein de l'équipe SIG**, différents travaux, qui peuvent faire l'objet de nouvelles thèses, doivent être poursuivis dans les domaines de la modélisation, l'interopérabilité et la personnalisation en RI notamment. Dans le même temps, vu que ces travaux s'intéressent de plus prêt à l'usager, les collaborations avec des chercheurs dans les sciences connexes (telles que la psychologie, les sciences cognitives) doivent s'amplifier. Ces collaborations peuvent prendre différentes formes telles que des **projets** (projets européens par exemple). Ces projets seraient en effet l'opportunité de rassembler des chercheurs ayant tous le même objectif et ce dans des champs d'applications différents (par exemple le « dossier médical patient », la recherche d'information). Différents champs d'applications pourraient alors être envisagés tel que celui des **systèmes ambiants**.

	Réponses apportées		
Hétérogénéité	**Fédération des activités documentaires**	**Amélioration des SRI**	**Modélisation pour la personnalisation en RI**
Des besoins informationnels	* Systèmes de recommandation	* Système de médiation pour la désambiguisation la RI. * Plateforme de RI à base de composants	* Modèle de ressource flexible et interopérable
Des caractéristiques des usagers de ces systèmes	* Processus adaptatifs	* Plateforme d'évaluation des interfaces	
L'environnement dans lequel la recherche d'information se déroule		* Interface multi-facettes * Plateforme d'évaluation des interfaces	* Modèle de contexte
Des activités en relation avec la recherche d'information	* Architecture basée sur des annotations collectives	* Plateforme de RI à base de composants	

Tableau 9 - Réponses que nous avons apportées aux problèmes d'hétérogénéités

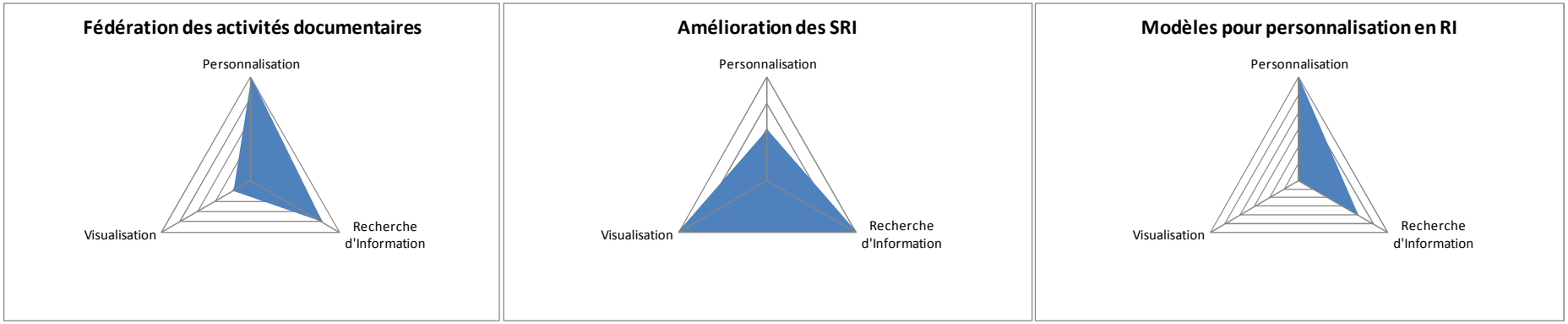


Figure 52 - Implication des principaux domaines de recherche dans nos travaux

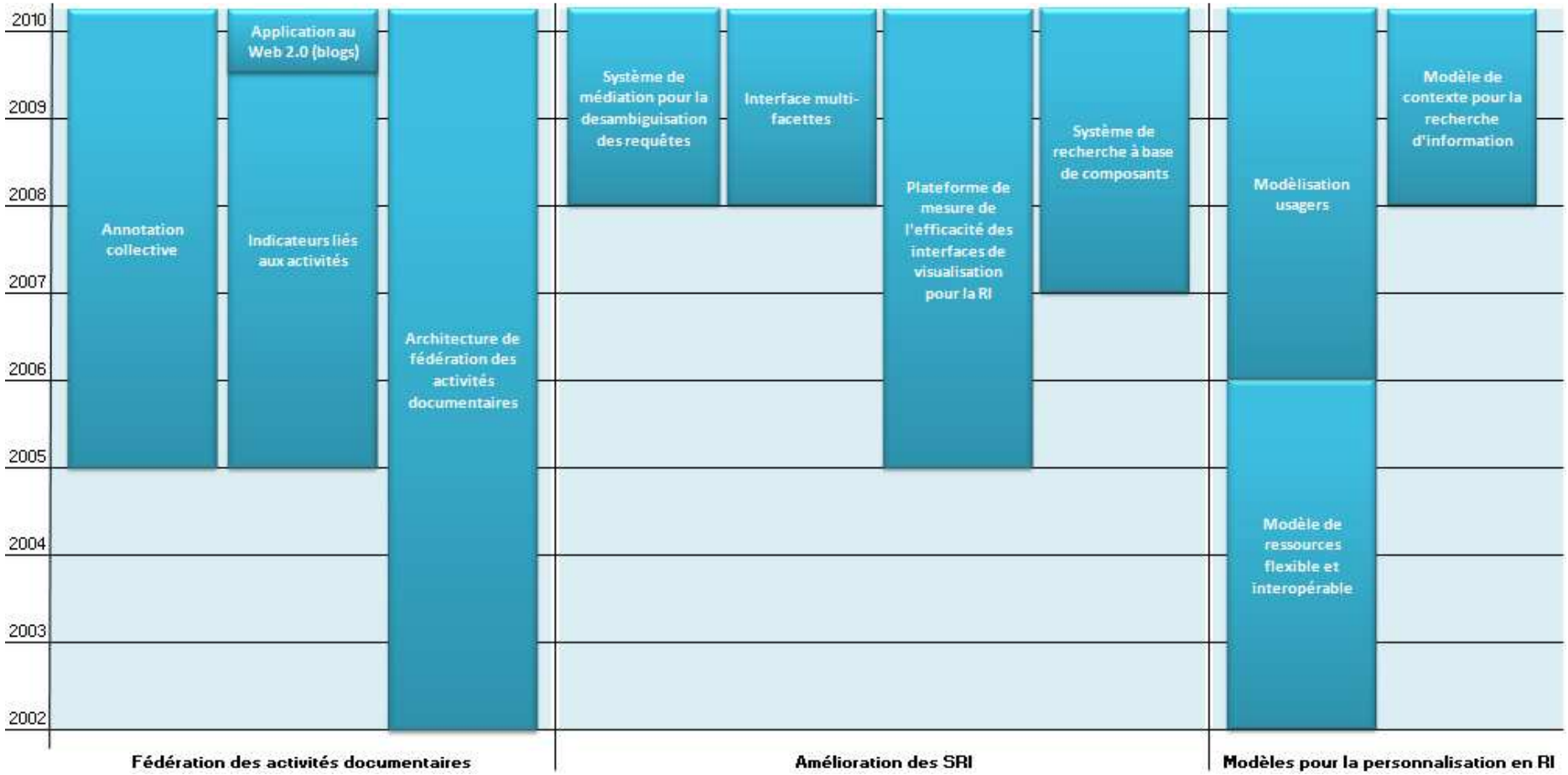


Figure 53 – Structuration temporelle des travaux

H. BIBLIOGRAPHIE

Dans cette liste, les références précédées par ☆ précisent celles dans lesquelles je fais partie des auteurs. La liste des auteurs, selon l'usage dans notre équipe, sont classés par ordre alphabétique.

Abrams, D., Baecker, R. et Chignell, M. 1998. Information Archiving with Bookmarks: Personal Web Space Construction and Organization. [éd.] ACM Press. *CHI'98: Proceedings of the conference on Human factors in computing systems.* 1998, pp. 41-48.

Adler, M.J. et van Doren, C. 1972. *How to read a book.* [éd.] Simon & Shuster. New York : s.n., 1972.

Agosti, M. et Ferro, N. 2005. Annotations as Context for Searching Documents. *CoLIS '05: Proceedings of the 5th International Conference on Conceptions of Library and Information Sciences.* 2005, LNCS 3507, pp. 155-170.

Agosti, M. et Smeaton, A. 1996. *Information Retrieval and Hypertext.* s.l. : Kluwer Academic Publisher, 1996. ISBN 0-7923-9710-X.

Agrawal, R. et Srikant, R. 1994. Fast algorithms for mining association rules. *Proc. 20th Int. Conference on Very Large Data Bases, VLDB94.* 1994.

Airio, E., Järvelin, K., Suomela, S., Saatsi, P. et Kekäläinen, J. 2004. CIRI - An Ontology-based Query Interface for Text Retrieval. *The Web Intelligence symposium.* 2004.

Amadieu, F., Tricot, A. et Mariné, C. 2009. Effects of prior knowledge diversity on learning with a non-linear electronic document: disorientation and coherence of the reading sequence. *Computers in Human Behavior.* 2009, Vol. 25, pp. 381-388.

Ardissono, L., Console, L. et Torre, I. 2001. An adaptive system for the personalised access to news. *AI Communications.* 2001, Vol. 14, pp. 129-147.

Arezki, R., Poncelet, P., Dray, G. et Pearson, D. W. 2004. Web Information Retrieval Based on User Profile. *Adaptive Hypermedia and Adaptive Web-Based Systems.* 2004, Vol. LNCS 3137, pp. 275-278.

Armstrong, R., Freitag, D. et Joachims, T. 1995. Webwatcher: machine learning and hypertext. *In Proceedings of the 1995 AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments.* 1995.

Baeza-Yates, R. et Ribeiro-Neto, B. 1999. *Modern Information Retrieval.* s.l. : Addison-Wesley Educational Publishers Inc, 1999. ISBN 978-0201398298.

Balfe, E. et Smyth, B. 2005. An Analysis of Query Similarity in Collaborative Web Search. *27th European Conference on IR Research (ECIR 2005): Advances in Information Retrieval.* 2005, Vol. LNCS 3408, pp. 330-344.

Barrett, R., Maglio, P. P. et Kellem, D. C. 1997. How to personalize the web. *International ACM Conference on Human Factors and Computing Systems (CHI).* 1997, pp. 75-82.

Bazire, M. et Brézillon, P. 2005. Understanding Context Before Using It. *Modeling and Using Context.* 2005, pp. 29-40.

Belbèze, C. et Soulé-Dupuy, C. 2007. Apport des services Web dans l'amélioration de l'accès à l'information sur le Web ? *Conférence francophone en Recherche d'Information et Applications (CORIA).* 2007.

☆ **Belbèze, C., Chevalier, M. et Soulé-Dupuy, C. 2009b.** Agrégats de mots-clés validés sémantiquement : pour de nouveaux services d'accès à l'information sur Internet. *Document numérique.* 2009b, Vol. 12, 1, pp. 81-105.

☆ **Belbèze, C., Chevalier, M. et Soulé-Dupuy, C. 2009c.** Nouveaux services basés sur des agrégations de mots-clés validées sémantiquement. *INFormatique des Organisations et Systemes d'Information et de Decision (INFORSID 2009).* 2009c, pp. 161-176.

☆ **Belbèze, C., Chevalier, M. et Soulé-Dupuy, C. 2009.** Semantic Comparison of Keywords Aggregates. *IADIS International Conference - Information Systems.* 2009, pp. 161-168.

☆ **Belbèze, C., Chevalier, M. et Soulé-Dupuy, C. 2008.** Web Services Based Information Access Architecture. *IADIS International Conference WWW/Internet (ICWI 2008).* 2008, pp. 119-127.

Belkin, N. et Croft, W. 1992. Information Retrieval and Information Filtering: Two Sides of the same Coin. *Communications of the ACM.* 1992, Vol. 35, 12, pp. 29-38.

Berisha-Bohe, S. et Rumpler, B. 2007. Modèle évolutif d'un profil utilisateur : Application à la Recherche d'Information dans une bibliothèque numérique de thèses. *Conférence francophone en Recherche d'Information et Applications.* 2007, pp. 197-209.

Biancalana, C. 2009. Social Tagging for Personalized Web Search. *AI*IA 2009: Emergent Perspectives in Artificial Intelligence.* 2009, pp. 232-242.

☆ **Bonnel, N. et Chevalier, M. 2006.** Critères d'évaluation pour les interfaces des systèmes de recherche d'information. *Conférence francophone en Recherche d'Information et Applications.* 2006, pp. 109-114.

☆ **Bonnel, N. et Chevalier, M. 2006b.** Evaluation des interfaces utilisateur d'information. *Atelier Visualisation et extraction de connaissances - Conférence EGC 2006.* 2006b.

Bonnel, N. 2006. Génération dynamique de présentations interactives en multimédia 3D, de données, pour des applications en ligne. *Thèse de Doctorat, Université de Rennes 1.* 2006.

☆ **Bonnel, N., Chevalier, M. et Dousset, B. 2008.** Métaphores de visualisation des résultats de recherche d'information sur le Web. *Recherche d'information: état des lieux et perspectives.* Recherche d'information et web - dirigée par Bernadette Bouchon-Meunier, 2008, pp. 295-339.

☆ **Bonnel, N., Chevalier, M., Dousset, B. et Hubert, G. 2008b.** Visualisation en recherche d'information. [éd.] Sophie Chauvin. *Information & visualisation - enjeux, recherches et applications.* 2008b, pp. 165-187.

Bottraud, J.-C. et Bisson, G. 2004. Expansion de requêtes par apprentissage automatique dans un assistant pour la recherche d'information. *1ère conférence francophone en Recherche d'Information et Applications (CORIA).* 2004, pp. 89-107.

Boughanem, M. et Tmar, M. 2001. Filtrage d'information par combinaison d'un profil positif et un profil négatif. *Third Colloquium of the French ISKO Chapter: Filtrage et résumé automatique de l'information sur les réseaux.* 2001, pp. 209-217.

Bouzeghoub, M. et Kostadinov, D. 2005. Personnalisation de l'information: aperçu de l'état de l'art et définition d'un modèle flexible de profils. *Conférence francophone en Recherche d'information et Applications (CORIA).* 9-11 mars 2005, pp. 201-218.

☆ **Boyer, M., Canut, M.-F., Chevalier, M., Péninou, A. et Sèdes, F. 2007.** Cartographie de l'organisation : une approche topologique. *Journées Francophones Extraction et Gestion de Connaissances (EGC 2007),.* 2007, Vol. 2, pp. 557-568.

Brajnik, G. et Tasso, C. 1994. A Shell for Developing Non-monotonic User Modeling Systems. *International Journal of Human-Computer Studies.* 1994, Vol. 40, pp. 31-62.

Bringay, S., Barry, C. et Charlet, J. 2004. Les documents et les annotations du dossier patient hospitalier. *Information - Interaction - Intelligence.* 2004, Vol. 4, 1, pp. 191-211.

Broder, A., Ciccolo, P., Gabrilovich, E. et Pang, B. 2008. Domain-Specific Query Augmentation using Folksonomy Tags: the Case of Contextual Advertising. *Proceedings of the 1st Workshop on Information Retrieval for Advertising, in conjunction with the 31st ACM International Conference on Research and Development in Information Retrieval.* 2008, pp. 23-26.

Brookes, B.C. 1997. The developping cognitive viewpoint in information science. [éd.] M. and al. De Mey. 1997, pp. 195-203.

Brusilovsky, P. et Millán, E. 2007. User Models for Adaptive Hypermedia and Adaptive Educational Systems. *The Adaptive Web.* 2007, pp. 3-53.

Byström, K. et Hansen, P. 2005. Conceptual framework for tasks in information studies: Book Reviews. *J. Am. Soc. Inf. Sci. Technol.* 2005, Vol. 56, pp. 1050-1061.

Cabanac, G. 2002. Interface de Classification de Signets Web. *Rapport de stage de l'IUT Informatique Toulouse.* 2002.

Cabanac, G. 2008b. Annotation collective dans le contexte RI : définition d'une plate-forme pour expérimenter la validation sociale. *CORIA/RJCRI'08 : 3e Rencontres Jeunes Chercheurs en Recherche d'Information.* 2008b, pp. 385-392.

Cabanac, G. 2008. *Fédération et amélioration des activités documentaires par la pratique d'annotation collective.* Toulouse, Université Toulouse 3 : s.n., 2008. Thèse de Doctorat, spécialité informatique, Université Toulouse 3.

☆ **Cabanac, G., Chevalier, M., Chrisment, C et Julien, C. 2006d.** L'architecture CoMED pour la gestion collective de documents électroniques dans l'organisation. *CIDE'06 : 9e Colloque International sur le Document Électronique.* 2006d, pp. 237-252.

☆ **Cabanac, G., Chevalier, M., Chrisment, C. et Julien, C. 2007.** Valoriser et intégrer les activités documentaires de l'organisation grâce à l'annotation collective de documents électroniques. *VSST'07 : actes du 5e colloque Veille Stratégique, Scientifique & Technologique.* 2007.

☆ **Cabanac, G., Chevalier, M., Chrisment, C. et Julien, C. 2005.** A Social Validation of Collaborative Annotations on Digital Documents. *Proceedings of the International Workshop on Annotation for Collaboration - Methods, Tools and Practices.* 2005, pp. 31-40.

☆ **Cabanac, G., Chevalier, M., Chrisment, C. et Julien, C. 2009.** Activités documentaires des usagers au sein de l'organisation : amélioration par la pratique d'annotation collective. *Ingénierie des Systèmes d'Information.* 2009, Vol. 14, 3, pp. 97-117.

☆ **Cabanac, G., Chevalier, M., Chrisment, C. et Julien, C. 2007b.** An Original Usage-based Metrics for Building a Unified View of Corporate Documents. *DEXA'07: Proceedings of the 18th International Conference on Database and Expert Systems Applications.* 2007b, pp. 202-212.

☆ **Cabanac, G., Chevalier, M., Chrisment, C. et Julien, C. 2007c.** Collective Annotation: Perspectives for Information Retrieval Improvement. *RIAO'07: Proceedings of the 8th conference on Information Retrieval and its Applications.* 2007c.

☆ **Cabanac, G., Chevalier, M., Chrisment, C. et Julien, C. 2008.** Exploiting the Annotation Practice for Personal and Collective Information Management. *International Workshop on Model Driven Information Systems Engineering: Enterprise, User and System Models (CAiSE/MoDISE-EUS 2008).* 2008, Vol. 341, pp. 67-78.

☆ **Cabanac, G., Chevalier, M., Chrisment, C. et Julien, C. 2008b.** Exploiting the Annotation Practice for Personal and Collective Information Management. *INFORSID/PeCUSI'08 : 2e atelier Prise en Compte de l'Usager dans les Systèmes d'Information.* 2008b, pp. 55-66.

☆ **Cabanac, G., Chevalier, M., Chrisment, C. et Julien, C. 2010.** Social validation of collective annotations: Definition and experiment. [éd.] Wiley. *Journal of American Society for Information Science and Technology.* 2010, Vol. 61, 2, pp. 271-287.

☆ **Cabanac, G., Chevalier, M., Chrisment, C. et Julien, C. 2006.** Validation sociale d'annotations collectives : argumentation bipolaire graduelle pour la théorie sociale de l'information. *INFORSID'06 : 24e congrès de l'INFormatique des Organisations et Systèmes d'Information et de Décision.* 2006, pp. 467-482.

☆ **Cabanac, G., Chevalier, M., Chrisment, C. et Julien, C. 2009c.** Visualisation et exploration du capital documentaire d'une organisation au travers d'une interface multifacette. *Ingénierie des Systèmes d'Information.* 2009c. Vol. 14, 2, pp. 35-60.

☆ **Cabanac, G., Chevalier, M., Chrisment, C., Julien, C., Soulé-Dupuy, C. et Tchienehom, P. 2008c.** Web Information Retrieval: Towards Social Information Search Assistants. *Social Information Technology: Connecting Society and Cultural Issues.* 2008c, pp. 218-252.

☆ **Cabanac, G., Chevalier, M., Ravat, F. et Teste, O. 2007d.** An Annotation Management System for Multidimensional Databases. *DaWaK'07: Proceedings of the 9th International Conference on Data Warehousing and Knowledge Discovery.* 2007d, pp. 89-98.

☆ **Cabanac, G., Chevalier, M., Ravat, F. et Teste, O. 2009b.** Decisional Annotations: Integrating and Preserving Decision-Makers' Expertise in Multidimensional Systems. *Complex Data Warehousing and Knowledge Discovery for Advanced Retrieval Development: Innovative Methods and Applications.* Advances in Data Warehousing and Mining, 2009b, p. Chapitre 4.

☆ **Cabanac, G., Chevalier, M., Ravat, F. et Teste, O. 2006b.** Méta-modélisation des bases de données multidimensionnelles annotées. *Revue des Nouvelles Technologies de l'Information (RNTI-B-2) - Entrepôts de Données et Analyse en ligne (EDA'06).* 2006b, pp. 39-54.

☆ **Cabanac, G., Chevalier, M., Ravat, F. et Teste, O. 2006c.** Modèle conceptuel pour bases de données multidimensionnelles annotées. *Revue des Nouvelles Technologies de l'Information : EGC.* 2006c, pp. 119-124.

Cadiz, J. J., Gupta, A. et Grudin, J. 2000. Using Web Annotations for Asynchronous Collaboration Around Documents. *CSCW'00: Proceedings of the 2000 ACM conference on Computer supported cooperative work.* 2000, pp. 309-318.

☆ **Canut, M.-F., Chevalier, M., Gunia, N., Mottay, D. et Péninou, A. 2005b.** Du SIRH Transactionnel au SIRH Décisionnel : exploitation Dds Sociogrammes pour la Gestion des Compétences. *Colloque de l'AIM (Association Information et Management) (AIM 2005).* 2005b.

☆ **Canut, M.-F., Chevalier, M., Gunia, N., Mottay, D. et Péninou, A. 2005.** DU SIRH TRANSACTIONNEL AU SIRH DECISIONNEL : EXPLOITATION DES SOCIO-GRAMMES POUR LA GESTION DES COMPETENCES. *Colloque de l'AIM (Association Information et Management) (AIM 2005).* 2005.

☆ **Canut, M.-F., Chevalier, M., Péninou, A. et Sèdes, F. 2004.** Cartographie de l'organisation : une représentation basée sur les acteurs et leurs relations. *séminaire Veille Stratégique Scientifique & Technologique (VSST'2004).* 2004.

☆ **Canut, M.-F., Chevalier, M., Péninou, A. et Sèdes, F. 2005.** Modélisation des individus et de leurs relations pour l'aide à l'intégration des individus dans l'organisation. *Revue des Nouvelles Technologies de l'Information - Numéro spécial Actes des Journées francophones d'Extraction et Gestion des Connaissances - EGC 2005.* 2005, Vol. E-3, 2, pp. 599-610.

Carman, M. J., Baillie, M. et Crestani, F. 2008. Tag Data and Personalized Information Retrieval. *ACM Workshop on Search in Social Media SSM'08.* 2008, pp. 27-34.

Cayrol, C. et Lagasquie-Schiex, M.-C. 2005. Gradual Valuation for Bipolar Argumentation Frameworks. *ECSQARU'05: Proceedings of the 8th European Conference on Symbolic and Qualitative Approaches to Reasoning and Uncertainty.* 2005, pp. 366-377.

☆ **Chaker, H., Chevalier, M., Soulé-Dupuy, C. et Tricot, A. 2010b.** Adapter le contexte métier pour améliorer le système de recherche d'information . *Prise en Compte de l'Usager dans les Systèmes d'Information (INFORSID) (Atelier PeCUSI 2010).* 2010b.

☆ **Chaker, H., Chevalier, M., Soulé-Dupuy, C. et Tricot, A. 2010c.** Improving information retieval by modelling business context . *International Workshop on User Profiles in Multi-application Environments (at CENTRIC 2010) (MultiA-Pro 2010.* 2010c.

☆ **Chaker, H., Chevalier, M., Soulé-Dupuy, C. et Tricot, A. 2010.** Système de recherche d'information pour les tâches métier. *Conférence francophone en Recherche d'Information et Applications (CORIA 2010).* 2010, pp. 301-312.

Chakrabarti, S. 2007. Dynamic Personalized Pagerank in Entity Relation Graphs. *WWW2007.* 2007, pp. 571-580.

Chen, C. 2006. Information Visualization - Beyond the Horizon. [éd.] Springer London LTD. 2006.

☆ **Chevalier, M. et Ciaccia, A. 2009.** Modélisation utilisateurs et systèmes d'information sur le web : pour une approche centrée usager. *Hypertextes, Hypermédias (H2PTM 2009).* 2009, pp. 61-73.

☆ **Chevalier, M. et Hubert, G. 2005.** Evaluation d'une interface de restitution de résultats de recherche : Quelles conclusions en tirer ? *Atelier Visualisation des Connaissances - Conference EGC 2005.* 2005.

☆ **Chevalier, M. et Hubert, G. 2009c.** Query Assistant based on Experience Capitalization for Information Retrieval Systems. *IEEE International Conference on Human System Interaction.* 2009c, pp. 502-509.

☆ **Chevalier, M. et Julien, C. 2002b.** Aide à la navigation sur le Web. *Revue Extraction des Connaissances et Apprentissage - Actes des journées EGC.* 2002b, Vol. 1, 4, pp. 399-406.

☆ **Chevalier, M. et Julien, C. 2003.** Interface adaptative et coopérative pour l'aide à la Recherche d'information sur le Web. *Information - Interaction - Intelligence.* 2003, Vol. 3, 2, pp. 47-73.

☆ **Chevalier, M. et Julien, C. 2003b.** Interface coopérative et adaptative pour la Recherche d'Information sur le Web. *19èmes Journées de Bases de Données Avancées.* 2003b.

☆ **Chevalier, M. 2002.** *Interface adaptative pour l'aide à la recherche d'information sur le Web.* Toulouse : IRIT, 2002. Thèse de Doctorat, spécialité Informatique, Université Toulouse III. Soutenue le 16 décembre 2002.

☆ **Chevalier, M., Chrisment, C. et C., Julien. 2004.** Helping people searching the web: towards an adaptive and a social system. *IADIS International Conference WWW/Internet (ICWI 2004.* 2004, pp. 405-412.

☆ **Chevalier, M., Hubert, G. et Mothe, J. 2006.** GenIE (Genomic Information Extraction). *Rapport de recherche.* 2006.

☆ **Chevalier, M., Julien, C et Soulé-Dupuy, C. 2008.** Collaborative and Social Information Retrieval and Access: Techniques for improved User Modeling. [éd.] C. Julien, C. Soulé-Dupuy M. Chevalier. 2008.

☆ **Chevalier, M., Julien, C. et Khrouf, K. 2003.** Towards a Documentary Memory: Building a Document Repository for Companies. *International Conference on Enterprise Information Systems (ICEIS'03).* 2003, pp. 213-218.

☆ **Chevalier, M., Julien, C. et Soulé-Dupuy, C. 2008b.** Profils usagers pour la recherche d'information : pertinence de leur usage ? *Atelier PeCUSI - Congrès Inforsid* . 2008b.

☆ **Chevalier, M., Julien, C., Soulé-Dupuy, C. et Valles-Parlangeau, N. 2007.** Personalized Information Access Through Flexible and Interoperable Profiles. *International Web Information Systems Engineering - International Workshop on Personalized Access to Web Information (WISE-PAWI 2007).* 2007, pp. 374-385.

☆ **Chevalier, M., Julien, C., Valles-Parlangeau, N. et Soulé-Dupuy, C. 2009b.** Usagers et recherche d'information. *Ingénierie des Systèmes d'Information.* 2009b, Vol. 14, 3.

☆ **Chevalier, M., Soulé-Dupuy, C. et Tchienehom, P. 2004.** A profile-based architecture for a flexible and personalized information access. *IADIS/WWW Internet 2004.* 2004.

☆ **Chevalier, M., Soulé-Dupuy, C. et Tchienehom, P. 2006b.** Interopérabilité de profils pour l'accès à des ressources. *Congrès Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID 2006).* 2006b, pp. 65-80.

☆ **Chevalier, M., Soulé-Dupuy, C. et Tchienehom, P. 2005b.** Profiles re-usability for personalized information access: application to users contexts determination. *nternational Workshop on Context-based Information Retrieval (CIR'05), in conjunction with Context'05.* 2005b, pp. 71-82.

☆ **Chevalier, M., Soulé-Dupuy, C. et Tchienehom, P. 2005c.** Profiles Semantics and Matchings Flexibility for Ressources Access. *IEEE international Conference on Signal Image Technology and Internet based Systems.* 2005c, pp. 224-231.

☆ **Chevalier, M., Soulé-Dupuy, C. et Tchienehom, P. 2006.** Semantics-based Profiles Modeling and Matching for Resources Access. *Journal des Sciences pour l'Ingénieur.* 2006, Vol. 1, 7, pp. 54-63.

Chi, E. H. 2000. A Taxonomy of Visualization Techniques Using the Data State Reference Model. *Proceedings of the IEEE Symposium on Information Visualization (InfoVis).* 2000, pp. 69-75.

Cho, Y. H., Kyeong, J. et Kim, S.H. 2002. A personalized recommender system based on web usage mining and decision tree induction. *Expert System with Applications.* 2002, Vol. 23, 3, pp. 329-342.

Christine, M. et Lainé-Cruzel, S. 1999. Profil-Doc : un prototype de système de recherche d'information personnalisé selon le profil des utilisateurs. *11ème Conférence Francophone sur l'Interaction Homme-Machine.* 1999.

Ciaccia, A. 2008. *De la compréhension d'un besoin d'information à la recherche d'information sur le Web : analyse cognitive du rôle de facteurs liés à l'utilisateur et à la tâche.* 2008. Thèse de doctorat, Université Paris-Ouest Nanterre.

Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* 1960, Vol. 20, 37, pp. 37-46.

Cool, C. et Spink, A. 2002. Issues of context in information retrieval (IR): an introduction to the special issue. *Information Processing & Management.* 2002, Vol. 38, pp. 605-611.

Cornuéjols, A. et Miclet, L. 2002. *Apprentissage Artificiel.* s.l. : Eyrolles, 2002. ISBN 2-212-11020-0.

Danilowicz, C. et Nguyen, H. C. 2002. Using User Profiles in Intelligent Information Retrieval. *13th international Symposium on Foundations of intelligent Systems.* 2002, Vol. LNCS 2366, pp. 223-231.

Daoud, M., Tamine-Lechani, L. et Boughanem, M. 2009. Towards a graph based user profile modeling for a session-based personalized search. *Knowledge and Information Systems.* 2009, Vol. 21, 3, pp. 365-398.

Denoue, L. 2000. *De la création à la capitalisation des annotations dans un espace personnel d'informations.* 2000. Thèse de doctorat, Université de Savoie, France.

DeRose, S., Daniel, R., Grosso, P., Maler, E., Marsh, J. et Walsh, N. 2002. XML Pointer Language (XPointer). *W3C.* 2002.

Dupont, C et Watson, M. 2010. Stars make searcj more personal. *The official Google blog.* [En ligne] 03 mars 2010. [Citation : 20 juin 2010.] http://googleblog.blogspot.com/2010/03/stars-make-search-more-personal.html.

Egyed-Zsigmond, E., Mille, A. et Prié, Y. 2003. Club ♣ (Trèfle): A Use Trace Model. *International conference on case-based reasoning (ICCBR).* 2003, pp. 146-160.

Ehang, J. et Ackerman, M.S. 2005. "Searching For Expertise in Social Networks: A Simulation of Potential Strategies. [éd.] ACM Press. *Group'05: Proceedings of the international conference on Supporting Group Work.* 2005, pp. 71-80.

Ertzscheid, O. et Gallezot, G. 2003. *Chercher faux et trouver juste, sérendipité et recherche d'informations.* Bucarest : X° Colloque bilatéral franco-roumain, CIFSIC, – Atelier D2 - "Communication et complexité", 2003. URL : http://archivesic.ccsd.cnrs.fr/documents/archives0/00/00/06/89/sic_00000689_02/sic_00000689.html.

Fekete, J.D. et Plaisant, C. 2004. Les leçons tirées de deux compétitions de visualisation d'information. *16th Conference on Association Francophone d'interaction Homme-Machine (IHM 2004).* 2004, pp. 7-12.

Feldman, S. 2004. The high cost of not finding information. *KM World magazine.* 2004, Vol. 13, 3.

Ferreira, J., Da Silva, A. R. et Delgado, J. 2005. Web Services for Information Retrieval. *IEEE – International Conference on Information Technology: Coding and Computing.* 2005, Vol. 2, 4-6, pp. 497-502.

Finin, T. W. 1989. GUMS: A general user modeling shell. [éd.] Springer-Verlag. *User Models in Dialog Systems.* 1989, pp. 411-430.

Fink, J et Kobsa, A. 2000. A Review and Analysis of Commercial User Modeling Servers for Personalization on the World Wide Web. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research.* 2000, Vol. 10, pp. 209-249.

Fleiss, J. L. 1971. Measuring Nominal Scale Agreement Among Many Raters. *Psychol. Bull.* 1971, Vol. 76, 5, pp. 378-382.

Foley, C., Smeaton, A. et Jones, G. J. F. 2008. Combining Relevance Information in a Synchronous Collaborative Information Retrieval Environment. [éd.] Christine Julien, Chantal Soulé-Dupuy Max Chevalier. *Collaborative and Social Information Retrieval and Access: Techniques for improved user modeling.* 2008, pp. 140-164.

Fortunato, S. 2010. Community detection in graphs. *Physics Reports, Vol. 486, No. 3-5. (February 2010), pp. 75-174.* 2010, Vol. 486, 3-5, pp. 75-174.

Fraenkel, A. S. et Klein, S. T. 1999. Information Retrieval from Annotated Texts. *J. Am. Soc. Inf. Sci.* 1999, Vol. 50, 10, pp. 845-854.

Furnas, G.W. et Zacks, J. 1994. Multitrees: enriching and reusing hierarchical structure. *ACM International Conference on Human Factors in Computing Systems (CHI'94).* 1994, pp. 330-336.

Goecks, J. et Mynatt, E. D. 2004. Leveraging social networks for information sharing. *Proceedings of the 2004 ACM conference on Computer supported cooperative work.* 2004, pp. 328-331.

Goker, A. et McCluskey, T. L. 1991. Towards an Adaptive Information Retrieval System. *6th international Symposium on Methodologies For intelligent Systems.* 1991, Vol. LNCS 542, pp. 348-357.

Golder, S.A. et Huberman, B.A. 2006. Usage patterns of collaborative tagging systems. *Journal of Information Science.* 2006, Vol. 32, 2, pp. 198-208.

Gong, B., Peng, B. et Li, X. 2007. A Personalized Re-ranking Algorithm Based on Relevance Feedback. *Advances in Web and Network Technologies, and Information Management.* 2007, pp. 255-263.

Gwizdka, J. et Lopatovska, I. 2009. The role of subjective factors in the information search process. *Journal of the American Society for Information Science and Technology.* 2009, Vol. 60, 12, pp. 2452-2464.

Hammond, T., Hannay, T., Lund, B et Scott, J. 2005. Social Bookmarking Tools (I): A General Review. *D-Lib Magazine.* 2005, Vol. 11, 4.

Harb, A., Dray, G., Plantié, M., Poncelet, P., M., Roche et Trousset, F. 2008. Détection d'Opinions : Apprenons les bons Adjectifs ! *Actes de l'Atelier FOuille des Données d'OPinions (FODOP 08), conjointement à la Conférence INFORSID 08.* 2008.

Harper, D. J., Mechkour, M. et Muresan, G. 1999. Document Clustering for Mediated Information Access. *Proceedings of IRSG99: the 21st BCS-IRSG Annual Colloquium on IR Research.* 1999, pp. 256-268.

Hascoët, M. et Beaudouin-Lafon, M. 2001. Visualisation interactive d'information. *Revue Information-Interaction-Intelligence (I3).* 2001, Vol. 1, 1, pp. 77-108.

Hoang, H. H., Nguyen, T. M. et Tjoa, A M. 2008. A semantic Web Based Approach for context-aware User Query formulation and information retrieval. [éd.] Christine Julien, Chantal Soulé-Dupuy Max Chevalier. *Collaborative and Social Information Retrieval and Access: Techniques for improved user modeling.* Décembre 2008.

Hoashi, K., Kazunori, M., Naomi, I. et Hashimoto, K. 2000. Document filtering method using non relevant information profile. *In Proceedings of the twenty third Annual International ACM SIGIR Conference on Research and Development in Information Retrieval: Distributed Retrieval.* 2000, pp. 176-183.

Hoffman, C., Lomonosov, A. et Sitharam, M. 1997. Finding Solvable Subsets of Constraint Graphs. *International Conference on Principles and Practice of Constraint Programming.* 1997, pp. 436-477.

Hölscher, C. et Strube, G. 2000. Web search behavior of Internet experts and newbies. *9th International Conference on the World Wide Web (WWW9).* 15-19 may 2000.

Hubert, G. et Mothe, J. 2007. Reusing Past Queries to Facilitate Information Retrieval. [éd.] INSTICC Press. *2nd International Conference on Software and Data Technologies (ICSOFT 2007).* 2007, pp. 166-171.

Huete, J. F., de Campos, L. M., Fernandez-Luna, J. M. et Rueda-Morales, M. A. 2008. Hierarchical naive bayes models for representing user profiles. *Proceedings of the 31st Annual international ACM SIGIR Conference on Research and Development in information Retrieval.* 2008, pp. 711-712.

Hull, D. 1993. Using Statistical Testing in the Evaluation of Retrieval Experiments. *SIGIR'93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval.* 1993, pp. 329-338.

Humbert, J.-C. 2005. LA GESTION ÉLECTRONIQUE DOCUMENTAIRE (GED) DANS LES DISCIPLINES MÉDICALES ET SCIENTIFIQUES. *Bulletin de l'Académie Lorraine des Sciences.* 2005, Vol. 44, 1-4.

Ihadjadène, M. 1999. *La recherche et la navigation dans un système de recherche d'information grand public : le cas des hypercatalogues sur l'Internet.* Lyon : s.n., 1999. Thèse de Doctorat, spécialité Science de l'Information et de la Communication, Université Lyon 1. Soutenue le 17 décembre 1999.

Ingwersen, P. et Järvelin, K. 2005. The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series). [éd.] Springer-Verlag. 2005.

Jackson, H. J. 2002. *Marginalia: Readers writing in books.* s.l. : Yale University Press, 2002.

Jaczynski, M. et Trousse, B. 1998. WWW Assisted Browsing by Reusing Past Navigations of a Group of Users. *4th European Workshop on Advances in Case-Based Reasoning (EWCBR).* 1998, Vol. LNCS 1488, pp. 160-171.

Jansen, B. J., Spink, A. et Saracevic, T. 2000. Real life, real users and real needs: a study and analysis of user queries on the web. *Information Processing and Management.* 2000, Vol. 36, pp. 207-227.

Jardine, N. et van Rijsbergen, C. J. 1971. The Use of Hierarchic Clustering in Information Retrieval. *Information Storage and Retrieval.* 1971, Vol. 7, 5, pp. 217-240.

Jeon, H., Kim, T. et Choi, J. 2008. Adaptive User Profiling for Personalized Information Retrieval. *3rd International Conference on Convergence and Hybrid Information Technology.* 2008, Vol. 2, pp. 836-841.

Jéribi, L. et Rumpler, B. 2002. Instance Cooperative Memory to Improve Query Expansion in Information Retrieval Systems. *Journal of Universal Computer Science.* 2002, Vol. 8, 6, pp. 591-601.

Jermann, J. 2002. Résolution de contraintes géométriques par rigidification récursive et propagation d'intervalles. *Thèse de Doctorat, Université de Nice Sophia-Antipolis.* 2002.

Jing, Y. et Croft, W.B. 1994. An association thesaurus for information retrieval. *RIAO.* 1994, pp. 146-160.

Joachims, T. 2002. Optimizing search engines using clickthrough data. *8th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (KDD'02).* 2002, pp. 133-142.

Johnson, J. D. 2003. On contexts of information seeking. *Information Processing and Management.* 2003, Vol. 39, pp. 735-760.

Jones, W. 2007. How People Keep and Organize Personal Information. [éd.] University of Washington Press. *Personal information management.* 2007, pp. 35-56 (chap. 3).

Jones, W., Phuwanartnurak, A. J., Gill, R. et Bruce, H. 2005. Don't Take My Folders Away!: Organizing Personal Information to Get Things Done. [éd.] ACM Press. *CHI'05 extended abstracts on Human factors in computing systems.* 2005, pp. 1505-1508.

Juin, A. 2010. Application de recherche d'information personnalisée. *Rapport de Stage IUT Information, Rangueil, Toulouse III.* 2010.

Jung, S., Herlocker, J. L. et Webster, J. 2007. Click data as implicit relevance feedback in web search. *Information Processing and Management.* 2007, Vol. 4, 3, pp. 791-807.

Kahan, J., Koivunen, M.-R., Prud'Hommeaux, E et wick, R. R. 2002. Annotea: an open RDF infrastructure for shared Web annotations. *Comp. Netw.* 2002, Vol. 32, 5, pp. 589-608.

Kahan, J., Koivunen, M.-R., Prud'Hommeaux, E. et Swick, R. R. 2002. Annotea: an open RDF infrastructure for shared Web annotations. *Comp. Netw.* 2002, Vol. 32, 5, pp. 589-608.

Kautz, H. A., Selman, B. et Shah, M. A. 1997. Referral Web: Combining Social Networks and Collaborative Filtering. *Communications of the ACM.* 1997, Vol. 40, 3, pp. 63-65.

Kay, J. 1994. Lies, damned lies and stereotypes: pragmatic approximations of users. *Proc. of Fourth International Conference on User Modeling.* 1994, pp. 175-184.

Kay, J. 1995. The um Toolkit for reusable, long term user models. *User Modeling and User-Adapted Interaction.* 1995, Vol. 4, 3, pp. 149-196.

Kay, J., Kummerfeld, B. et Lauder, P. 2002. Personis: A Server for User Models. *Adaptive Hypermedia and Adaptive Web-Based Systems: Second International Conference.* 2002, pp. 203-212.

Kaye, J. J., Vertesi, J., Avery, S., Dafoe, A., David, S., Onaga, L., Rosero, I. et Pinch, T. 2006. To Have and to Hold: Exploring the Personal Archive. *CHI'06: Proceedings of the conference on Human Factors in computing systems.* 2006, pp. 275-284.

Kelly, D. et Teevan, J. 2003. Implicit Feedback for inferring user preference : a bibliography. *SIGIR Forum.* 2003, Vol. 37, 2.

Khoo, C. S., Luyt, B., Ee, C., Osman, J., Lim, H.-H. et Yong, S. 2007. How users organize electronic files on their workstations in the office environment: a preliminary study of personal information organization behaviour. *Information Research.* janvier 2007, Vol. 11, 2.

Kidd, A. 1994. The marks are on the knowledge worker. [éd.] ACM Press. *CHI'94: Proceedings of the SIGCHI conference on Human factors in computing systems.* 1994, pp. 186-191.

Kirchhoff, L., Stanoevska-Slabeva, K., Nicolai, T. et Fleck, M. 2008. Using social network analysis to enhance information retrieval systems. *5th Conference on Applications of Social Network Analysis (ASNA).* 2008.

Klink, S. 2004. Improving Document Transformation Techniques with Collaborative Learned Term-Based Concepts. *Reading and Learning: adaptive content recognition.* 2004, Vol. LNCS 2956, pp. 281-305.

Kobsa, A. et Fink, J. 2006. An LDAP-Based User Modeling Server and its Evaluation. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research.* 2006, Vol. 16, pp. 129-169.

Kobsa, A. 2007. Generic User Modelling Systems. *Teh adaptive Web.* 2007, pp. 136-154.

Kobsa, A. 2007. Privacy-Enhanced Web Personalization. [éd.] A. Kobsa, and W. Nejdl P. Brusilovsky. *The Adaptive Web.* 2007, pp. 628-670.

Kohonen, T., Kaski, S., Lagus, K., Honkela, J., Paatero, V. et Saarela, A. 2000. Self Organization of a Massive Document Collection. *IEEE Transaction on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery.* 2000, Vol. 11, 3, pp. 574-585.

Kompaore, N. D. et Mothe, J. 2008. Fusion de résultats en recherche d'information : Mesure de l'impact de l'union et de l'intersection de résultats. *Conférence Internationale sur le Document Electronique (CIDE 2008).* 2008, pp. 3-15.

Korfhage, R. R. 1997. *Information storage and retrieval.* [éd.] Wiley. 1997. ISBN 978-0471143383 .

Koychev, I. 2000. Gradual Forgetting for Adaptation to Concept Drift. *ECAI 2000 Workshop Current Issues in Spatio-Temporal Reasoning.* 2000.

Kules, B., Capra, R., Banta, M. et Sierra, T. 2009. What do exploratory searchers look at in a faceted search interface? *JCDL'09: Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries.* 2009, pp. 313-322.

Kwasnik, B. H. 1999. The role of classification in knowledge representation and discovery. *Library Trends.* 1999, Vol. 8, 2-3, pp. 22-47.

Laflaquière, J. Settouti, L. Prié, Y. et Mille, A. 2006. Trace-based framework for Experience Management and Engineering. *Lecture Notes in Computer Science.* 2006, Vol. 4251, pp. 1171-1178.

Lainé-Cruzel, S. 1999. ProfilDoc – Filtrer une information exploitable. *Bulletin des bibliothèques de France (BBF).* 1999, Vol. 44, 5, pp. 60-64.

Li, C. 2007. Forrester's new Social Technographics report. *Groundswell.* [En ligne] 23 avril 2007. [Citation : 17 juin 2010.] http://forrester.typepad.com/groundswell/2007/04/forresters_new_.html.

Li, Y. et Belkin, N. J. 2008. A faceted approach to conceptualizing tasks in information seeking. *Information Processing and Management.* 2008, Vol. 44, pp. 1822-1837.

Lieberman, H. 1995. Letizia: an agent that assists web browsing. *International Joint Conference on Artificial Intelligence (IJCAI'95).* 1995.

Liu, B. 1998. *Web Data Mining.* s.l. : Springer, 1998. ISBN 978-3-540-37881-5.

Luce, R. D. et Perry, A. D. 1949. A Method of matrix analysis of group structure. *Psychometrika.* 1949, Vol. 14, pp. 95-116.

Lund, B., Hammonf, T, Flack, M. et Hannay, T. 2005. Social bookmarking tool (II): A case study - Connotea. *D-Lib Magazine.* 2005, Vol. 11, 4.

Maarek, Y. S. et Ben-Shaul, I. 1996. Automatically Organizing Bookmarks per Contents. *Computer Networks and ISDN Systems.* 1996, Vol. 28, 7-11, pp. 1321-1333.

Malone, T. W., Grant, K. R., Turbak, A., Brobst, S. A. et Cohen, M. D. 1987. Intelligent information sharing systems. *Communication of the ACM.* 1987, Vol. 30, 5, pp. 390-402.

Manning, C.D., Raghavan, P. et Shütze, H. 2008. *Introduction to Information Retrieval.* s.l. : Cambridge University Pres, 2008. ISBN 0521865719.

Marchionini, G. 1997. *Information seeking in electronic environments.* [éd.] Cambridge University Press. 1997. Vol. Issue 9 of Cambridge series on human-computer interaction. ISBN 9780521586740 .

Markov, Z. et Larose, D. T. 2007. *Data Mining the Web.* s.l. : Wiley, 2007. ISBN 978-0-471-66655-4.

Marshall, C.C. 1998. Toward an Ecology of Hypertext Annotation. *Proceedings of the 9th conference on Hypertext and hypermedia.* 1998, pp. 40-49.

Martins, D. S., Biajiz, M., do Prado, A. F. et de Souza, W. L. 2009. Implicit relevance feedback for context-aware information retrieval in UbiLearning environments. *ACM Symposium on Applied Computing (SAC'09).* 2009, pp. 659-663.

Mechkour, , M., Harper, D. J. et Muresan, G. 1998. The WebCluster project. Using clustering for mediating access to the World Wide Web. *21st annual international conference ACM SIGIR.* 1998, pp. 357-358.

Meyer zu Eissen, S. et Stein, B.,. 2007. An MDA Approach to Implement Personal IR Tools. *4th International Workshop on Text-based Information Retrieval in conjunction with DEXA 2007.* 2007.

Millen, D.R., Feinberg, J. et Kerr, B. 2006. Dogear: Social bookmarking in the enterprise. [éd.] ACM Press. *CHI'06: proceeding of the ACM SIGCHI Conference on Human Factors in Computing Systems.* 2006, pp. 111-120.

Mizzaro, S. et Tasso, C. 2002. Personalization techniques in the tips project: The cognitive filtering module and the information retrieval assistant. *Proceedings of the Workshop on Personalization Techniques in Electronic Publishing on the Web: Trends and Perspectives.* 2002.

Mizzaro, S. 2007. Relevance, the whole (hi) story. *Journal of the American society for information science.* 2007, Vol. 48, 9, pp. 810-832.

Molinaro, F. 2004. Document et information. *Les infostrateges.com.* [En ligne] 16 janvier 2004. [Citation : 10 juin 2010.] http://www.les-infostrateges.com/article/040171/document-et-information.

Montaner, M., Lopez, B. et de la Rosa, J.L. 2003. A taxonomy of recommender agents on the Internet. *Artificial Intelligence Review.* 2003, Vol. 19, 4, pp. 285-330.

Mylonas, P., Vallet, D., Castells, P., Fernández, M. et Avrithis, Y. 2008. Personalized information retrieval based on context and ontological knowledge. *Knowledge Engineering Review.* 2008, Vol. 23, 1, pp. 73-100.

Nanas, N., Uren, V. et De Roeck, A. 2003. Building and applying a concept hierarchy representation of a user profile. *26th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR'03).* 2003, pp. 198-204.

Nguyen, A.-T., Denos, N. et Berrut, C. 2006. Modèle d'espaces de communautés basé sur la théorie des ensembles d'approximation dans un système de filtrage hybride. *Conférence francophone en Recherche d'Information et Application CORIA.* 2006, pp. 303-314.

Ovsiannikov, I. A., Arbib, M. A. et McNeill, T. H. 1999. Annotation technology. *Int. J. Hum.-Comput. Stud.* 1999, Vol. 50, 4, pp. 329-362.

Paterno, F. 2000. Model-baseddesign of interactive application. *intelligence.* 2000, Vol. 11, 4, pp. 26-38.

Pazzani, M., Muramatsu, J. et Billsus, D. 1996. Syskill & Webert: Identifying interesting web sites. *National Conference on Artificial Intelligence.* 1996, pp. 54-61.

Pédauque, R. T. 2006. Document et modernités. 2006.

Picard, R. W. 1997. Affective Computing. [éd.] MIT Press. 1997.

Plaisant, C., Fekete, J. D. et Grinstein, G. 2008. Promoting Insight Based Evaluation of Visualizations: From Contest to Benchmark Repository. *IEEE Transactions on Visualization and Computer Graphics.* 2008, Vol. 14, 1, pp. 120-134.

Port, B. 2009. Développement d'une plateforme de test pour interfaces graphiques dédiées à la recherche d'information. *Rapport de stage L3 ISI, Toulouse.* 2009.

Pretschner, A. et Gauch, S. 1999. Ontology Based Personalized Search. *11th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 99).* 1999, pp. 391-398.

Quinlan, J.R. 1993. C4.5: programs for machine learning. 1993.

Quiroga, L. M. et Mostafa, J. 1999. Empirical evaluation of explicit versus implicit acquisition of user profiles in information filtering systems. *Proceedings of the fourth ACM conference on Digital libraries .* 1999, pp. 238-239.

Raghavan, V. V. et Sever, H. 1995. On the reuse of past optimal queries. *18th Annual Internation ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'95).* 1995, pp. 344-350.

Ravat, F. et Teste, O. 2008. Personalization and OLAP Databases. *Annals of Information Systems, Springer, S. Kozielski, R. Wrembel*, Vol. 3, 2008.

Reiterer, H., Mann, T. M. et Mubler, G. 2001. Visual Information Retrieval for the WWW. *Proc. Human Computer Interaction (HCI).* 2001, pp. 1150-1154.

Renner, S. A. 2001. A "Community of Interest" approach to data interoperability. *Federal Database Colloquium.* 2001.

Rocchio, J. J. 1971. Relevance Feedback in Information Retrieval. [éd.] Prentice Hall Inc. *The SMART Retrieval System - Experiments in Automatic Document Processing.* 1971, pp. 313-323.

Rosenfeld, L. et Morville, P. 1998. Information Architecture for the World Wide Web. [éd.] Inc O'Reilly & Associates. First Edition, 1998.

Rücker, J. et Polanco, M.J. 1997. Siteseer: personalized navigation for the Web. *Communications of the ACM.* 1997, Vol. 40, 3, pp. 73-76.

Ruthven, I. et Lalmas, M. 2003. A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review.* 2003, Vol. 18, 2, pp. 95-145.

Salton, G. 1971. A comparison between manual and automatic indexing methods. *Journal of the American Documentation.* 1971, Vol. 20, 1, pp. 61-71.

Salton, G. et Buckley, C. 1990. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science.* 1990, Vol. 41, 4, pp. 288-297.

Sanderson, M. 2008. Ambiguous queries: test collections need more sense. *SIGIR 2008: Proceedings of the 31st annual international ACM SIGIR conference on Research and Decelopment in information retrieval.* 2008, pp. 499–506.

Savy, N. 2006. Probabilités et statistiques pour modéliser et décider : tests, validation, régression, plans d'expérience. *Statistiques.* Ellipses, 2006.

Schilit, B.N., Golovchinsky, G. et Price, M.N. 1998. Beyond Paper: Supporting Active Reading with Free Form Digital Ink Annotations. *CHI 98 conference summary on Human factors in computing systems.* 1998, pp. 249-256.

Schreck, J. 2001. *Security and Privacy in User Modeling.* Universitat – Gesamthochschule – Essen : s.n., 2001. Doctorat. URL : http://www.security-and-privacy-in-user-modeling.info/.

Sellen, A. J. et Harper, R.H. 2003. *The myth of the paperless office.* Cambridge, MA, USA : MIT Press, 2003.

Shapira, B., Shoval, P. et Hanani, U. 1997. Stereotypes Information Filtering Systems. *Information Processing & Management.* 1997, Vol. 33, 3, pp. 273-287.

Shingo, O. et Masaru, K.,. 2006. Clustering of Search Engine Keywords Using Access Logs. *Conference on Database and Expert Systems Applications, DEXA 2006.* 2006, pp. 842-852.

Sieg, A., Mobasher, B. et Burke, R. 2007. Web search personalization with ontological user profiles. *6th ACM Conference on Conference on information and Knowledge Management (CIKM'07).* 2007, pp. 525-534.

Silverstein, C., Henzinger, M., Marais, H. et Moricz, M. 1998. Analysis of a very large web search engine query log. *SRC technical note #1998-014.* 26 octobre 1998.

Somlo, G. L. et Howe, A. E. 2003. Using web helper agent profiles in query generation. *2nd international Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'03).* 2003, pp. 812-818.

Soulé-Dupuy, C. 2001. Bases d'informations textuelles : des modèles aux applications. *Habilitation à diriger des recherches, Université Paul Sabatier.* 2001.

Speretta, M. et Gauch, S. 2005. Personalized search based on user search histories. *IEEE/WIC/ACM International Conference on Web Intelligence.* 2005, pp. 622-628.

Spink, A., Jansen, B. J. et Wolfram, D., Saracevic, T. 2002. From e-sex to e-commerce: web search changes. *Revue IEEE Computer.* mars 2002, Vol. 35, 3.

Stan, J., Egyed-Zsigmond, E., Joly, A. et Maret, P. 2008. A User Profile Ontology For Situation-Aware Social Networking. *3rd Workshop on Artificial Intelligence Techniques for Ambient Intelligence (AITAmI2008).* 2008.

Stefani, A. et Strappavara, C. 1998. Personalizing Access to Web Sites: The SiteIF Project. *2nd Workshop on Adaptive Hypertext and Hypermedia HYPERTEXT'98.* 1998.

Sugiyama, K., Hatano, K. et Yoshikawa, M. 2004. Adaptive web search based on user profile constructed without any effort from users. *Proceedings 13 th International Conference on World Wide Web.* 2004, pp. 675-684.

Swarts, J. 2004. Cooperative Writing: Achieving Coordination Together and Apart. [éd.] ACM Press. *SIGDOC'04: proceedings of the 22th annual international conference on Design of Communication.* 2004, pp. 83-89.

Tamine, L. 2008. De la recherche d'information orientée système à la recherche d'information orientée contexte : Verrous, contributions et perspectives. *Habilitation à diriger des recherches, Université Paul Sabatier.* 2008.

Tamine, L., Boughanem, M. et Zemirli, W. 2007. Exploiting Multi-Evidence from Multiple User's Interests to Personalizing Information Retrieval. *IEEE International Conference on Digital Information Management(ICDIM 2007).* 2007, pp. 7-12.

Teste, O. 2000. Elaboration d'entrepôts de données complexes. *XVIIIème congrès INFormatique des ORganisations et Systèmes d'Information et de Décision (INFORSID'00)*, Lyon (France), mai 2000, pp. 229-245.

Tomasic, A., Simmons, I. et Zimmerman, J. 2007. Learning Information Intent via Observation. 2007.

Trousse, B., Jaczynski, M. et Kanawati, R. 1999. Using User Behavior Similarity for Recommendation Computation: The Broadway Approach. *8th international conference on human computer interactions (HCI'99).* 1999, pp. 85-89.

Tufféry, S. 2010. *Data Mining et statistiques décisionnelle.* [éd.] Technip. 2010. ISBN 978-2-7108-0946-3.

Turney, P. 2002. Thumbs up or thumbs down ? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of 40th Meeting of the Association for Computational Linguistics.* 2002, pp. 417-424.

Usunier, N., Amini, M. et Gallinari, P. 2006. Résumé automatique de texte avec un algorithme d'ordonnancement. *Revue Ingénierie des Systèmes d'Information.* Hermès, 2006, Vol. 11, 2, pp. 71-91.

Vergara, H. 1994. PROTUM: A Prolog Based Tool for User Modeling. [éd.] University of Konstanz Department of Information Science. *WG Knowledge-Based Information Systems.* 1994.

Vinay, V., Wood, K., Milic-Frayling, N. et Cox, I. J. 2005. Comparing relevance feedback algorithms for web search. *14th international Conference on World Wide Web - Special interest Tracks and Posters, WWW'05.* 2005, pp. 1052-1053.

Viviani, M., Bennani, N. et Egyed-Zsigmond, E. 2010. A Survey on User Modeling in Multi-application Environments. *User Profiles in Multi-application Environments In conjunction with CENTRIC 2010, August 22-27, 2010 - Nice, France.* 2010.

Voorhees, E. M. et Harman, D. 1999. Overview of the Eighth Text REtrieval Conference (TREC-8). *8th Text REtrieval Conference.* 1999.

Voorhees, E. M. 2001. Overview of TREC 2001. *TREC'01: Proceedings of the 10th Text REtrieval Conference.* 2001.

Wexelblat, A. et Maes, P. 1997. Footprints: History-Rich Web Browsing. *Conférence Internationale en Recherche d'Information Assistée par Ordinateur (RIAO).* 1997, pp. 75-85.

White, R., Ruthven, I. et Jose, J. M. 2002. The Use of Implicit Evidence for Relevance Feedback in Web Retrieval. *24th BCS-IRSG European Colloquium on IR Research: Advances in information Retrieval.* 2002, Vol. LNCS 2291, pp. 93-109.

Widmer, G. et Kubat, M. 1996. Learning in the presence of concept drift and hidden contexts. [éd.] Kluwer Academic Publisher. *Machine Learning.* 1996, Vol. 23, pp. 69-101.

Widyantoro, D. H. 1999. Dynamic Modeling and Learning User Profile in Personal News Agen. *Master Thesis, Dept. of Computer Science.* 1999.

Wilcoxon, F. 1945. Individual comparisons by ranking methods. *Biom. Bull.* 1945, Vol. 1, 6, pp. 80-83.

Wolfe, J. L. 2000. Effects of Annotations on Student Readers and Writers. *Proceedings of the 5th ACM conference on Digital libraries.* 2000, pp. 19-26.

Zamir, O. et Etzioni, O. 1999. Grouper: A Dynamic Clustering Interface to Web Search Results. *Computer Networks.* 1999, Vol. 31, 11-16, pp. 1361-1374.

Zamir, O. 1998. Visualisation of search results in document retrieval systems. *General Examination, University of Washington.* 1998.

Zhuang, Z. et Cucerzan, S. 2006. Re-ranking search results using query logs. *15th ACM international Conference on information and Knowledge Management (CIKM '06).* 2006, pp. 860-861.

I. ANNEXE 1 : ARBORESCENCE MESH

Cette liste correspond à l'arborescence MeSH utilisée dans nos expérimentations. Elle contient 146 nœuds et 4974 documents. Pour chaque catégorie, nous précision son identifiant, son nom, le nombre de nœuds fils précisé entre <> et le nombre de documents associés à ce nœud. Un document peut être associé à plusieurs catégories.

```
@@ 957 Cardiovascular Diseases <5> 0
-@@ 958 Cardiovascular Abnormalities <2> 0
--@@ 960 Arteriovenous Malformations <1> 32
---@@ 961 Arteriovenous Fistula <0> 47
--@@ 964 Heart Defects, Congenital <6> 148
---@@ 965 Aortic Coarctation <0> 44
---@@ 968 Coronary Vessel Anomalies <0> 29
---@@ 971 Ductus Arteriosus, Patent <0> 23
---@@ 974 Heart Septal Defects <2> 19
----@@ 977 Heart Septal Defects, Atrial <0> 30
----@@ 980 Heart Septal Defects, Ventricular <0> 29
---@@ 983 Tetralogy of Fallot <0> 25
---@@ 984 Transposition of Great Vessels <0> 39
-@@ 990 Heart Diseases <13> 0
--@@ 991 Arrhythmia <8> 299
---@@ 993 Atrial Fibrillation <0> 62
---@@ 994 Atrial Flutter <0> 13
---@@ 995 Bradycardia <0> 40
---@@ 999 Heart Block <1> 85
----@@ 1001 Bundle-Branch Block <0> 20
---@@ 1003 Long QT Syndrome <0> 12
---@@ 1005 Pre-Excitation Syndromes <1> 0
----@@ 1008 Wolff-Parkinson-White Syndrome <0> 39
---@@ 1009 Sick Sinus Syndrome <0> 9
---@@ 1010 Tachycardia <2> 217
----@@ 1011 Tachycardia, Paroxysmal <0> 14
----@@ 1012 Tachycardia, Supraventricular <1> 71
-----@@ 1014 Tachycardia, Atrioventricular Nodal Reentry <0> 15
--@@ 1025 Cardiac Output, Low <0> 10
--@@ 1027 Cardiomegaly <1> 0
---@@ 1028 Cardiomyopathy, Congestive <0> 64
--@@ 1031 Endocarditis <0> 25
--@@ 1034 Heart Aneurysm <0> 27
--@@ 1035 Heart Arrest <0> 134
--@@ 1066 Heart Rupture <1> 26
---@@ 1067 Heart Rupture, Post-Infarction <0> 12
--@@ 1069 Heart Valve Diseases <7> 0
---@@ 1070 Aortic Valve Insufficiency <0> 60
---@@ 1071 Aortic Valve Stenosis <1> 76
----@@ 1074 Aortic Stenosis, Subvalvular <1> 0
-----@@ 1075 Cardiomyopathy, Hypertrophic <0> 59
---@@ 1077 Heart Murmurs <0> 9
---@@ 1078 Heart Valve Prolapse <1> 0
----@@ 1080 Mitral Valve Prolapse <0> 65
---@@ 1082 Mitral Valve Insufficiency <0> 67
---@@ 1083 Mitral Valve Stenosis <0> 49
---@@ 1089 Tricuspid Valve Insufficiency <0> 19
--@@ 1091 Myocardial Diseases <1> 93
---@@ 1101 Myocarditis <0> 38
--@@ 1102 Myocardial Ischemia <2> 0
---@@ 1103 Coronary Disease <5> 0
```

```
----@@ 1104 Angina Pectoris <2> 226
-----@@ 1105 Angina Pectoris, Variant <0> 25
-----@@ 1106 Angina, Unstable <0> 43
----@@ 1108 Coronary Aneurysm <0> 19
----@@ 1109 Coronary Arteriosclerosis <0> 51
----@@ 1110 Coronary Thrombosis <0> 40
----@@ 1111 Coronary Vasospasm <0> 44
---@@ 1113 Myocardial Infarction <1> 0
----@@ 1115 Shock, Cardiogenic <0> 20
--@@ 1117 Pericardial Effusion <0> 40
--@@ 1118 Pericarditis <0> 30
--@@ 1124 Rheumatic Heart Disease <0> 21
-@@ 1129 Hyperemia <0> 15
-@@ 1130 Pregnancy Complications, Cardiovascular <0> 88
-@@ 1136 Vascular Diseases <22> 0
--@@ 1137 Aneurysm <4> 96
---@@ 1138 Aneurysm, Dissecting <0> 50
---@@ 1142 Aneurysm, Infected <0> 13
---@@ 1143 Aneurysm, Ruptured <1> 0
----@@ 1144 Aortic Rupture <0> 41
---@@ 1145 Aortic Aneurysm <0> 107
--@@ 1155 Angiomatosis <1> 0
---@@ 1157 Hippel-Lindau Disease <0> 5
--@@ 1160 Angioneurotic Edema <0> 32
--@@ 1161 Aortic Diseases <1> 0
---@@ 1166 Aortic Arch Syndromes <0> 14
--@@ 1170 Arterial Occlusive Diseases <4> 180
---@@ 1171 Arteriosclerosis <1> 0
----@@ 1176 Intermittent Claudication <0> 25
---@@ 1178 Fibromuscular Dysplasia <0> 16
---@@ 1180 Mesenteric Vascular Occlusion <0> 25
---@@ 1182 Renal Artery Obstruction <0> 45
--@@ 1188 Arteritis <1> 13
---@@ 1193 Temporal Arteritis <0> 26
--@@ 1195 Cerebrovascular Disorders <6> 302
---@@ 1199 Carotid Artery Diseases <1> 155
----@@ 1200 Carotid Artery Thrombosis <0> 10
---@@ 1222 Cerebral Hemorrhage <0> 160
---@@ 1227 Cerebrovascular Accident <1> 0
----@@ 1228 Brain Infarction <1> 0
-----@@ 1231 Cerebral Infarction <0> 124
---@@ 1255 Intracranial Hemorrhages <2> 0
----@@ 1261 Intracranial Hemorrhage, Traumatic <2> 0
-----@@ 1265 Hematoma, Epidural <0> 13
-----@@ 1266 Hematoma, Subdural <0> 26
----@@ 1271 Subarachnoid Hemorrhage <0> 88
---@@ 1275 Vascular Headaches <2> 0
----@@ 1276 Cluster Headache <0> 12
----@@ 1277 Migraine <0> 110
---@@ 1287 Vertebrobasilar Insufficiency <0> 17
--@@ 1289 Diabetic Angiopathies <1> 0
---@@ 1291 Diabetic Retinopathy <0> 71
--@@ 1292 Embolism and Thrombosis <2> 0
---@@ 1293 Embolism <4> 44
----@@ 1294 Embolism, Air <0> 26
----@@ 1298 Embolism, Fat <0> 10
----@@ 1299 Pulmonary Embolism <0> 126
----@@ 1300 Thromboembolism <0> 76
---@@ 1310 Thrombosis <2> 254
----@@ 1312 Purpura, Thrombotic Thrombocytopenic <0> 10
----@@ 1323 Venous Thrombosis <3> 0
```

```
-----@@ 1324 Hepatic Vein Thrombosis <0> 20
-----@@ 1325 Retinal Vein Occlusion <0> 13
-----@@ 1326 Thrombophlebitis <0> 90
--@@ 1328 Hemorrhoids <0> 13
--@@ 1330 Hypertension <3> 0
---@@ 1333 Hypertension, Portal <1> 53
----@@ 1334 Esophageal and Gastric Varices <0> 75
---@@ 1335 Hypertension, Pulmonary <0> 83
---@@ 1337 Hypertension, Renal <1> 26
----@@ 1338 Hypertension, Renovascular <0> 69
--@@ 1340 Hypotension <1> 97
---@@ 1341 Hypotension, Orthostatic <0> 34
--@@ 1343 Ischemia <1> 202
---@@ 1346 Compartment Syndromes <0> 32
--@@ 1373 Raynaud's Disease <0> 31
--@@ 1381 Superior Vena Cava Syndrome <0> 11
--@@ 1382 Telangiectasis <1> 14
---@@ 1385 Telangiectasia, Hereditary Hemorrhagic <0> 9
--@@ 1386 Thoracic Outlet Syndrome <0> 4
--@@ 1388 Varicocele <0> 28
--@@ 1389 Varicose Veins <1> 0
---@@ 1390 Varicose Ulcer <0> 11
--@@ 1394 Vascular Hemostatic Disorders <6> 0
---@@ 1395 Cryoglobulinemia <0> 11
---@@ 1397 Ehlers-Danlos Syndrome <0> 11
---@@ 1398 Hemangioma, Cavernous <0> 24
---@@ 1400 Multiple Myeloma <0> 76
---@@ 1401 Osteogenesis Imperfecta <0> 13
---@@ 1404 Purpura, Schoenlein-Henoch <0> 15
--@@ 1410 Vasculitis <3> 75
---@@ 1418 Behcet's Syndrome <0> 16
---@@ 1420 Mucocutaneous Lymph Node Syndrome <0> 32
---@@ 1432 Wegener's Granulomatosis <0> 31
--@@ 1433 Venous Insufficiency <0> 20
```

Les caractéristiques de la sous-collection des maladies cardio-vasculaires sont résumées dans le Tableau 10.

Caractéristiques	Sous-collection
Nombre de documents	4974
Nombre de nœuds	146
Nombre de nœuds possédant au moins un document	119
Nombre de documents par nœud	Entre 4 et 302
Nombre total d'associations nœud/document	6417
Nombre moyen de documents par nœud ayant au moins un document	54
Proportion de nœuds ayant un nombre de documents inférieur ou égal à la moyenne	70 %
Longueur moyenne des documents	90 termes

Tableau 10 - Caractéristiques de la collection utilisée dans les expérimentations