Table des matières

Introduction générale

Contexte et Problématique

L'avènement de l'informatique, du multimédia et l'essor d'Internet sont, sans nul doute, l'une des plus grandes innovations de ce siècle. Les progrès des technologies de l'information, le large développement de nouveaux supports de média informatique (micro-ordinateur, téléphone portable, PDA) et l'amélioration des capacités de stockage sont les rouages essentiels de cette innovation. Les conséquences sur notre société ne sont pas des moindres. Notre perception et rapport avec l'information ont radicalement changé. L'expansion et la popularité grandissante du *web* ont en effet changé notre façon d'appréhender l'accès à la connaissance, la manière d'apprendre, de travailler et de vivre. Au delà des progrès et libertés apportés aux procédés d'éditions de l'information, cette innovation porte essentiellement sur l'ensemble des services de communication et d'accès à l'information. Elle a marqué le début d'une nouvelle ère de communication, d'une nouvelle société axée principalement autour de *l'information*.

Dès lors, l'accès à l'information est devenu un enjeu capital et stratégique : acquérir l'information pertinente, au bon moment, dès qu'elle est disponible est une nécessité pour tous, dans tous les domaines de la vie active. Un tel engouement a engendré une prolifération de nombreuses sources d'information. La quantité d'information disponible, particulièrement à travers le *web* n'a pas cessé d'augmenter. Nous sommes témoin d'une surabondance de l'information. Si on prend l'exemple du *web*, qui représente incontestablement la plus grande source d'information disponible jusqu'à présent et qui ne cesse de croître, un moteur de recherche populaire rapporte plus de huit (8) milliards de pages dans son index en juillet 2005 alors qu'elles étaient seulement 320 millions en 1997 et 3.3 milliards en septembre 2002. Le nombre d'utilisateurs est quand à lui estimé aujourd'hui à plusieurs centaines de millions. En conséquent, il devient de plus en plus difficile de retrouver précisément ce que l'on recherche dans cette masse de données.

L'élaboration de systèmes automatisés pour gérer ces masses de données est devenue dans un tel contexte une nécessité. La RI, domaine déjà ancien, est une branche en informatique qui s'intéresse à l'acquisition, l'organisation, le stockage et la recherche des informations. Elle propose des outils, appelés systèmes de recherche d'information (SRI), dont l'objectif est de capitaliser un volume important d'information et d'offrir des moyens permettant de

localiser les informations pertinentes relatives à un besoin en information d'un utilisateur exprimé à travers une requête.

Les efforts continus des chercheurs ont permis jusqu'à présent d'améliorer sans cesse les performances et la qualité des services d'accès à l'information. Les premiers travaux en RI qualifiés d'approche *classique*, sont apparus dans les années soixante, ils se sont focalisés à résoudre des problèmes principalement liés à la représentation de l'information, l'évaluation de requêtes ainsi que l'évaluation des performances de recherche. Toutefois, la RI *classique*, étant une approche orientée système, la différence de vocabulaire utilisé pour l'expression des contenus des documents et des besoins en information n'est pas prise en compte dans le processus d'appariement document-requête. Ce défaut d'appariement engendre une dégradation des performances de recherche. Ainsi, le problème qui se pose actuellement n'est plus tant la disponibilité de l'information mais la capacité d'accès et de sélection de l'information répondant aux besoins précis d'un utilisateur, à partir des représentations qu'il perçoit. Ces facteurs ont soulevé des défis majeurs pour les tâches de collecte et de gestion de l'information, le stockage, la transmission et la recherche efficace de l'information.

Dans un tel contexte, les travaux se sont orientés vers des approches dites *adaptatives* exploitant diverses sources d'évidence (documents jugés, termes pertinents, etc.) pour aider et assister l'utilisateur à retrouver les informations pertinentes à son besoin. Les travaux de la RI *adaptative* se sont particulièrement axés sur l'amélioration de l'efficacité du processus de recherche notamment lors de la phase d'exécution de la requête. Les techniques développées ont eu pour but de désambiguïser le sens des mots de la requête utilisateur afin de mieux cerner le but de sa recherche. Le principal objectif de ces approches est de générer une nouvelle requête plus adéquate que celle initialement formulée par l'utilisateur. Cette reformulation permet de coordonner le langage de recherche (utilisé par l'utilisateur dans sa requête) et le langage d'indexation des documents.

Néanmoins, nous pouvons constater à ce niveau un fossé entre le but de la RI et les méthodes qui la réalisent : l'objectif de la RI est de retrouver les documents pertinents répondant aux besoins en informations spécifiques de l'utilisateur alors qu'elle est implémentée de façon à ce que l'utilisateur soit uniquement représenté par sa requête. Même si elle n'est pas formulée ainsi, l'hypothèse suivante est implicitement faite dans les modèles adaptatifs : les résultats retournés pour une même requête sont identiques même si elle est exprimée par des utilisateurs différents. Cette hypothèse est clairement erronée de part la diversité des utilisateurs, de leurs besoins en information et de leurs difficultés à exprimer formellement leurs besoins par des mots clés. Elle est d'autant plus accentuée que la majorité des requêtes exprimées par les utilisateurs sont courtes et ambiguës, ce qui donne des spécifications inachevées sur leur besoin en informations. Lorsque l'utilisateur initie la recherche, c'est dans le but de combler un manque en information sur un sujet précis. Ce besoin en information est fortement lié à différents facteurs dépendants, d'une part, du but de sa recherche (son activité, son contexte et son environnement) et d'autre part, des connaissances capitalisées par l'utilisateur (ses préférences et centres d'intérêts) tout au long de ses expériences. Outre ce facteur humain, la majorité des moteurs de recherche disposent de peu de mécanismes permettant de s'adapter

au contexte précis d'un utilisateur dans le but d'augmenter la pertinence des informations retournées. Force est de constater qu'il n'y a généralement pas de mécanisme explicite qui représente et intègre l'utilisateur dans le processus de recherche adaptative.

De ce fait, au delà de la mise en œuvre des techniques d'*adaptation*, les travaux s'orientent actuellement vers la révision de la chaîne d'accès à l'information dans la perspective d'intégrer l'utilisateur dans l'ensemble des phases de recherche et ce, dans le but de lui délivrer l'information pertinente adaptée à son contexte et ses préférences, répondant à ses besoins précis. Dès lors l'accès à l'information tend vers une nouvelle définition [2] : "*Combine search technologies and knowledge about query and user context into a single framework in order to provide the most appropriate answer for a user's information needs*".

La personnalisation est un processus qui change la fonctionnalité, l'interface, la teneur en information, ou l'aspect d'un système pour augmenter sa pertinence personnelle en fonction des caractéristiques sociodémographiques déclarées de l'utilisateur (sexe, âge, lieu de résidence, etc.) et/ou de son comportement observé contenu dans ce que l'on nomme le *modèle* utilisateur. Ce modèle décrit toute information sur l'utilisateur, comme ses préférences, ses centres d'intérêts, ses besoins en information et son environnement de recherche. Les applications sont diverses : systèmes de recommandation, systèmes de filtrage de messages et d'informations, systèmes d'apprentissage, systèmes de recherche d'information (SRI), etc. Indépendamment de l'objectif applicatif visé, on identifie trois principaux aspects à promouvoir dans les systèmes d'accès personnalisé :

1. La capacité à identifier l'intention conceptuelle de l'utilisateur,
2. la flexibilité du processus de sélection de l'information en vue de s'adapter au contexte d'utilisation courant
3. et l'intelligibilité des interactions utilisateur - système.

Ainsi, l'efficacité de ces systèmes dépend fortement de la précision des modèles utilisateurs exploités. En effet, la modélisation du profil constitue l'élément essentiel dans le développement d'un système d'accès personnalisé performant. L'approche commune à tous les systèmes d'accès personnalisé à l'information, consiste en premier à modéliser le profil de l'utilisateur, puis à l'intégrer dans le processus d'accès à l'information. Cette procédure n'est cependant pas aussi simple, elle implique une représentation de l'utilisateur par un modèle ou une structure qui permettrait son exploitation par le SRI. En effet, la difficulté à ce niveau porte sur plusieurs points concernant l'étape de construction du profil utilisateur qui doit fournir une représentation fidèle des besoins et centres d'intérêts récurrents de l'utilisateur, ainsi que l'évolution de ce profil. D'autre part l'étape, d'exploitation du profil, dans le processus d'accès à l'information (reformulation de la requête, exécution de la requête, réordonnancement des résultats etc.), doit être en mesure d'intégrer les informations adéquates du profil permettant d'améliorer les résultats de recherche.

Les travaux présentés dans ce mémoire se situent dans le contexte précis de l'accès personnalisé à l'information. Plus particulièrement, cette thèse tente d'apporter des solutions à la

problématique majeure de la personnalisation de l'accès à l'information concernent principalement la modélisation et l'exploitation du profil de l'utilisateur.

Contributions

L'objectif de cette thèse est de proposer un nouveau modèle de RI personnalisé capable d'intégrer l'utilisateur dans le processus d'accès à l'information. Plus précisément, notre approche globale se focalise particulièrement sur deux aspects :

Le premier aspect porte sur la représentation et la maintenance d'une dimension qui caractérise les besoins récurrents de l'utilisateur en informations. Cette dimension qui est déterminée, évolue à partir de ses interactions avec le SRI, permet de définir ses centres d'intérêts.

Le second aspect concerne l'intégration de ces centres d'intérêts dans un modèle décisionnel d'accès à l'information. Notre objectif est de proposer un modèle inférentiel pour la mise en oeuvre du raisonnement lié à la prise de décision quant à la pertinence des documents compte tenu du profil de l'utilisateur d'une part et de la requête d'autre part. Cette fonction est fondée sur l'estimation d'une valeur de l'utilité de la décision prise quant à la sélection de documents pertinents. Comparativement aux travaux du domaine, notre approche s'en distingue fondamentalement par :

– l'intégration du profil de l'utilisateur comme composante à part entière du modèle formel d'accès personnalisé et non comme source de définition d'heuristiques ou techniques permettant la réécriture de la requête et/ou la fonction d'appariement,
– la caractérisation de l'accès personnalisé à l'information comme un problème décisionnel, ce qui justifie notre recours à la modélisation basée sur les diagrammes d'influence.

N'ayant pas de cadre standard d'évaluation, le second volet de notre contribution consiste en la définition d'un cadre d'évaluation adapté à l'accès personnalisé à l'information en augmentant les collections de la campagne TREC par des profils utilisateurs simulés. Pour évaluer le modèle que nous proposons nous avons besoin d'un modèle de référence permettant de quantifier l'apport du profil utilisateur dans le processus d'accès à l'information. Nous pouvons comparer notre modèle à n'importe quel modèle classique de recherche d'information ne tenant pas compte des centres d'intérêts de l'utilisateur. Cependant, notre modèle étant une extension des réseaux Bayésiens, il est plus significatif de considérer comme référence les résultats obtenus avec un tel modèle. Les résultats obtenus sont encourageants et ouvrent des perspectives intéressantes.

Organisation de la thèse

Ce mémoire est organisé en deux parties :

La première partie, regroupant quatre chapitres, présente un état de l'art sur la personnalisation de l'accès à l'information. Le *chapitre 1* présente les principaux facteurs d'émergence de la personnalisation dans le domaine de la RI, à travers l'évolution des SRIs classiques

vers les SRIs adaptatifs. Le *chapitre 2* introduit la thématique de recherche de la personnalisation de l'information à travers les éléments clés liés à l'entité *utilisateur* et présente la problématique générale de la personnalisation. Le *chapitre 3* traite de la théorie autour de la modélisation utilisateur. On y aborde les principales phases du processus de modélisation : les approches de représentation, les approches de construction et les méthodes d'évolution des profils utilisateurs. Le *chapitre 4* décrit les principaux modèles d'accès personnalisé à l'information et présente la problématique liée à la mise en place d'une campagne d'évaluation standard et formelle pour l'accès personnalisé ainsi qu'une synthèse des approches d'évaluation utilisées en RI personnalisée.

La seconde partie concerne nos contributions dans le domaine de la RI personnalisée. Nous donnons au début de cette partie les principales motivations pour la proposition d'un nouveau modèle d'accès personnalisé à l'information, ainsi que les grandes lignes de chacun des aspects de notre contribution en mettant en exergue sa spécificité relativement aux travaux du domaine. Puis, les trois chapitres de cette partie abordent en détail nos propositions. Le *chapitre 5* présente notre procédé de modélisation des dimensions informationnelles descriptives du profil utilisateur et l'outil développé pour l'acquisition implicite des documents visités par l'utilisateur. Le *chapitre 6* présente notre contribution à la formalisation d'un modèle d'accès personnalisé à l'information, via l'utilisation des diagrammes d'influence, extension des réseaux Bayésiens, intégrant le profil utilisateur proposé. Le *chapitre 7* présente le dernier volet de notre contribution qui consiste en la définition d'un cadre d'évaluation approprié pour l'accès personnalisé à l'information, que nous avons appliqué pour évaluer précisément notre modèle. Ce chapitre présente également les résultats expérimentaux obtenus lors des expérimentations portant essentiellement sur l'évaluation des performances et de l'impact des différents paramètres du modèle.

En conclusion, nous dressons un bilan de nos travaux, en mettant en exergue nos propositions. Nous présentons enfin les perspectives de nos travaux.

Le mémoire contient également trois annexes organisées comme suit :

L'*annexe A* présente un état de l'art portant sur la typologie des indicateurs d'intérêts pour l'observation du comportement et l'acquisition implicite des données utilisateurs lors de ses interactions avec le SRI. L'*annexe B* présente la description générale de l'outil développé *Web Cap*, implémentant l'algorithme d'inférence implicite proposé pour l'acquisition des données d'interaction de l'utilisateur lors de ses recherches. L'*annexe C* introduit les notions de base des réseaux Bayésiens, ainsi que les méthodes de calcul des probabilités conditionnelles dans ces réseaux.

Première partie

L’accès personnalisé à l’information

1 L'accès personnalisé à l'information : De la RI classique à la RI personnalisée

1.1 Introduction

La discipline de la recherche d'information (RI) est apparue dans un contexte où les progrès des technologies de l'information ont changé la perception de l'accès à l'information. Avec l'avènement du web, l'expansion de l'informatique à tous les domaines de la vie courante, a eu pour conséquence directe, l'accessibilité par un large public d'utilisateurs, autre que des documentalistes spécialisés, à des masses d'information volumineuses et hétérogènes.

Les efforts continus des chercheurs en RI ont permis jusqu'à présent d'améliorer sans cesse les performances et la qualité des services d'accès à l'information. Ce chapitre retrace les principales évolutions dans le domaine de la RI : de la première génération de systèmes de recherche d'information (SRI) dits *classiques* à la RI *adaptative*, puis récemment à la RI *personnalisée*.

La RI *classique*, apparue dans les années soixante, a une vision orientée système, en ce sens où la recherche des informations pertinentes se base uniquement sur l'appariement des documents avec la requête soumise par l'utilisateur. Toutefois, cette vision de l'accès à l'information suppose que l'utilisateur est extérieur au système de recherche. De plus, la difficulté qu'à l'utilisateur à exprimer son besoin en information par une requête, ainsi que la différence de vocabulaire entre les termes choisis par l'utilisateur pour formuler sa requête et les termes utilisés pour représenter les documents engendrent un défaut d'appariement. Ce défaut d'appariement est à l'origine d'une dégradation des performances de recherche. Cette problématique est encore plus accentuée avec l'accroissement continuel des sources d'information hétérogènes et la diversité des utilisateurs.

Dans un tel contexte, les travaux se sont orientés vers des approches dites *adaptatives* exploitant diverses sources d'évidence (documents jugés, termes pertinents, etc.) pour aider et assister l'utilisateur à retrouver les informations pertinentes à son besoin. Cependant, en dépit de l'efficacité des ces techniques adaptatives, le problème d'insatisfaction de l'utilisateur

persiste. On estime que 63% à 66% des 85% d'utilisateurs de moteurs de recherche sont insatisfaits aussi bien en termes de délai que de la qualité des réponses fournies [97].

Ainsi, dans le but de mieux répondre aux attentes et besoins des utilisateurs, les travaux en RI s'orientent actuellement vers des approches dites de *personnalisation* en exploitant des caractéristiques informationnelles spécifiques de l'utilisateur dans les processus d'accès à l'information.

Afin de mieux cerner cette évolution, nous discutons dans ce chapitre des principaux facteurs et problématiques ayants conduits à l'émergence de la RI personnalisée. Ce chapitre est organisé comme suit : la section 1.2 présente les fondements de la RI classique et donne un aperçu des principaux modèles. La section 1.3 présente les éléments précurseurs à l'émergence de la RI adaptative. La section 1.4 aborde les principales approches de la RI adaptative. La section 1.5, dresse un bilan de ses limites et présente les facteurs d'émergence de la RI personnalisée. Une conclusion générale sera présentée dans la dernière section.

1.2 Les fondements de la recherche d'information

L'objectif principal de la recherche d'information est de fournir des techniques et des outils pour sélectionner les informations pertinentes contenues dans une collection de documents en réponse aux besoins en information d'un utilisateur représentés à l'aide d'une requête. Nous citons ci-dessous la définition de la RI donnée par [187] dans sa forme originelle :

"The user expresses his information need in the form of a request for information. Information retrieval is concerned with retrieving those documents that are likely to be relevant to his information need as expressed by his request".

Cette définition fait apparaître deux notions clés que nous introduisons dans ce qui suit : document et requête utilisateur.

1.2.1 Notions de base

Document : Un document peut être un texte, un morceau de texte, une page web, une image, une vidéo, etc. On peut appeler document toute unité qui peut constituer une réponse à un besoin informationnel de l'utilisateur. Nous nous intéressons uniquement, dans ce travail, aux documents textuels. Dans la suite de cette thèse, nous utilisons indifféremment les termes *document* ou *information* pour désigner l'utilité documentaire retourné en réponse à la requête de l'utilisateur.

Requête : Une requête constitue l'expression du besoin en information de l'utilisateur. Elle représente l'interface entre le SRI et l'utilisateur. Divers types de langages d'interrogation sont proposés dans la littérature.

On peut citer :
– par une liste de mots clés : cas des systèmes SMART [149] et Okapi [143],
– en langage naturel : cas des systèmes SMART [149] et SPIRIT [61],
– en langage booléen : cas du système DIALOG [27],
– en langage graphique : cas du système NEURODOC [109].

1.2.2 Principales phases du processus de RI

L'objectif fondamental d'un processus de RI est de sélectionner les documents "*les plus proches*" du besoin en information de l'utilisateur décrit par une requête. Pour cela, le système de recherche regroupe un ensemble de méthodes et procédures permettant la gestion des collections de documents stockés sous forme d'une représentation intermédiaire permettant de refléter aussi fidèlement que possible leurs contenus sémantiques. L'interrogation de la collection de documents à l'aide d'une requête nécessite la représentation de cette dernière sous une forme unfiée compatible avec celles des documents. Ces fonctionnalités sont représentées à travers le processus global de la RI, communément nommé processus en U [14] et schématiquement illustré par la figure 1.1.

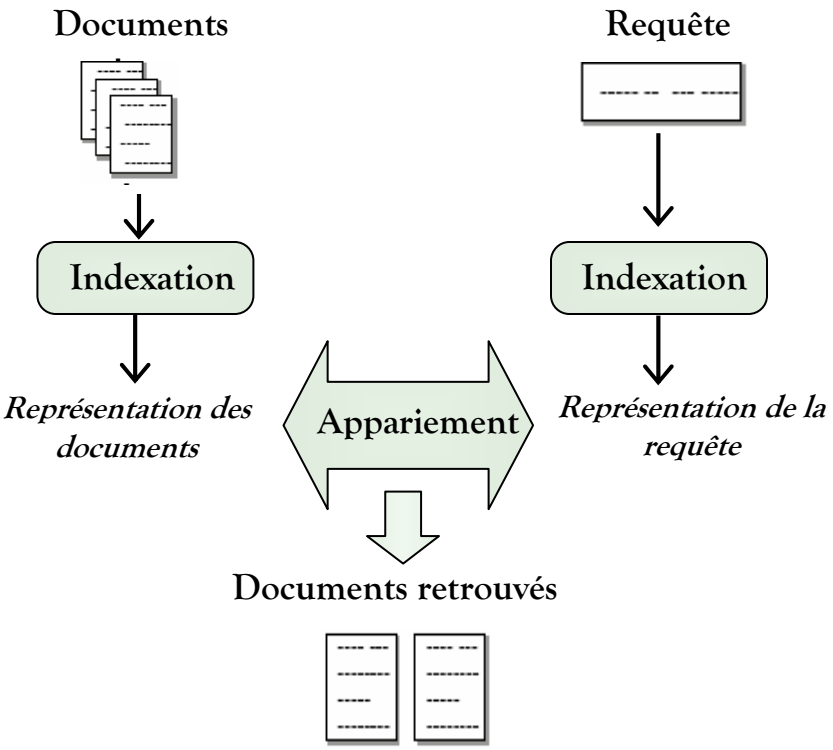


Figure 1.1 — Processus en U de la RI

Le déroulement de ce processus induit deux principales phases : indexation et appariement requête/document.

1.2.2.1 L'indexation

L'indexation est une étape très importante dans le processus de RI. Elle consiste à déterminer et à extraire les termes représentatifs du contenu d'un document ou d'une requête. La qualité de la recherche dépend en grande partie de la qualité de l'indexation. Le résultat de l'indexation constitue, ce que l'on nomme le **descripteur** du document ou de la requête. Ce dernier est souvent une liste de termes ou groupe de termes significatifs pour l'unité textuelle correspondante, généralement assortis de poids représentant leur degré de représentativité du contenu sémantique de l'unité qu'ils décrivent. Les descripteurs des documents (mots, groupe de mots) sont rangés dans un catalogue appelée dictionnaire constituant le **langage d'indexation**.

Techniquement, l'indexation peut être manuelle, automatique ou semi-automatique :

- *manuelle* : chaque document est analysé par un spécialiste du domaine ou un documentaliste.
- *automatique* : chaque document est analysé à l'aide d'un processus entièrement automatisé.
- *semi-automatique (mixte)* : c'est une combinaison des deux méthodes précédentes : un premier processus automatique permet d'extraire les termes du document. Cependant, le choix final reste au spécialiste du domaine ou au documentaliste pour établir les relations entre les mots clés et choisir les termes significatifs.

Les termes extraits des documents ne jouent pas le même rôle dans la représentation de ces derniers, en ce sens où ils n'ont pas le même degré d'importance. Pour caractériser ce degré de discrimination, il est courant en RI, d'affecter à chaque terme un poids. Cette étape est primordiable dans le processus d'indexation correspond au processus pondération. Pour trouver les termes du document qui représentent le mieux son contenu sémantique, [142] a défini la **fonction de pondération** d'un terme dans un document connue sous la forme de $TF * IDF$, qui est reprise dans différentes versions par la majorité des SRI [142; 152; 164; 161].

- TF (*term frequency*) : cette mesure est proportionnelle à la fréquence du terme dans le document. L'idée sous-jacente est que plus un terme est fréquent dans un document, plus il est important dans la description de ce document.
  Le TF est souvent exprimé selon l'une des déclinaisons suivantes :
  1. TF : utilisation brute,
  2. $\log(1+TF)$.
- IDF (*Inverse Document Frequency*) : mesure l'importance d'un terme dans toute la collection. L'idée sous-jacente est que les termes qui apparaissent dans peu de documents de la collection sont plus représentatifs du contenu de ces documents que ceux qui apparaissent dans tous les documents de la collection.

Cette mesure est exprimée selon l'une des déclinaisons suivantes :

1. $IDF = \log(\frac{N}{df})$,
2. $IDF = \log(\frac{N-df}{df})$

Où df est la proportion de documents contenant le terme et N le nombre total de documents dans la collection.

La fonction de pondération de la forme $TF * IDF$ consiste à multiplier les deux mesures TF et IDF. Une formule largement utilisée est la suivante :

$$TF * IDF = \log(1 + TF) * \log(\frac{N}{df}) \tag{1.1}$$

Une normalisation de la mesure du $TF * IDF$ par rapport à la longueur des documents a été proposée par [145; 161]. Une des formules les plus utilisées (citées) aujourd'hui dans le domaine de la RI est la formule $BM25$ d'OKAPI[145] tel que le poinds d'un terme i dans le document j (noté $w(i,j)$) est donnée par :

$$w(i,j) = 0.5 * \frac{tf_{ij} \star \log(\frac{N-n_i+0.5}{n_i+0.5})}{2 * (0.25 + \frac{0.75 * dl_j}{avgdl}) + tf_{ij}} \tag{1.2}$$

où : n_i : le nombre de documents contenant t_i,

N : le nombre de documents pertinents dans la collection,

dl : la longueur du document d_j,

$avg - dl$: la longueur moyenne des documents de la collection,

tf_{ij} : la fréquence d'apparition du terme t_i dans le document d_j

1.2.2.2 L'appariement document-requête

Le processus d'appariement document-requête permet de mesurer la pertinence d'un document vis-à-vis d'une requête. De manière générale, à chaque réception d'une requête, le SRI calcule un score de pertinence (similarité vectorielle, probabiliste, etc.). Ce score de pertinence est calculé à partir d'une fonction ou d'une mesure de similitude, notée $RSV(Q,D)$ (*Retrieval Status Value*) où Q est une requête et D un document de la collection. Le processus d'appariement est étroitement lié au processus d'indexation et de pondération des termes. Il existe deux méthodes d'appariement :

– *Appariement exact* (« *exact match retrieval* »)
  Le résultat est une liste de documents respectant exactement la requête spécifiée avec des critères précis. Les documents retournés ne sont pas triés [149].
– *Appariement approché* (« *best match retrieval* »)
  Le résultat est une liste de documents sensés être pertinents pour la requête. Les documents retournés sont triés selon leur score de pertinence vis-à-vis de la requête [142].

1.2.3 Taxonomie des modèles de RI

Les travaux de recherche dans le domaine de la RI ont conduit à la proposition de nombreux modèles [7]. Un modèle de RI a pour rôle de fournir une formalisation du processus de recherche d'information et un cadre théorique pour la modélisation de la mesure de pertinence. Nous présentons très brièvement dans ce qui suit les plus importants.

Modèle booléen [149]. Ce modèle propose la représentation d'une requête sous forme d'une expression logique. Les termes d'indexation sont reliés par les connecteurs logiques ET, OU et NON. Le processus de recherche mis en œuvre, consiste à effectuer des opérations sur les ensembles de documents définis par la présence et l'absence de termes d'indexation, afin de réaliser un appariement exact avec l'équation de la requête. Une extension de ce modèle a été effectuée par [151] : le modèle booléen étendu. Il intègre des poids dans l'expression de la requête et des documents. La sélection des documents s'effectuera donc sur la base d'un appariement rapproché et non plus exact.

Modèle vectoriel (Vector Space Model). Suivant la proposition faite part Luhn [116], le modèle vectoriel a été développé par Salton [149] [151] dans le projet SMART (*Salton's Magical Automatic Retriever of Text*). Ce modèle repose sur les bases mathématiques des espaces vectoriels. Les requêtes et les documents sont représentés dans l'espace vectoriel engendré par les termes d'indexation. Dans ce modèle, le degré de pertinence d'un document vis-à-vis de la requête est proportionnel à la position des deux vecteurs dans l'espace. Elle est évaluée à l'aide du degré de corrélation entre les vecteurs associés. Ce cœfficient de similarité (RSV) est calculé sur la base d'une fonction qui mesure la colinéarité des vecteurs documents et requête.

Modèle probabiliste (Probabilistic Model) [142; 145]. Ce modèle aborde la problème de la recherche d'information dans un cadre probabiliste. La pertinence document-requête est traduite par le calcul de la probabilité de pertinence d'un document par rapport à une requête. La pertinence entre un document et une requête est mesurée par le rapport entre la probabilité qu'un document D donné soit pertinent pour une requête Q, notée $p(R/D)$, et la probabilité qu'il soit non pertinent, notée $p(\overline{R}/D)$, ou R est l'évement de petinence et $\overline{R}$ de non pertinence. Ces probabilités sont estimées par les probabilités conditionnelles selon qu'un terme de la requête est présent, dans un document pertinent ou dans un document non pertinent.

1.3 De la RI classique à la RI adaptative

Le principe fondamental commun à tous les modèles classiques de RI suppose que les documents sélectionnés doivent contenir les mêmes mots (voir une partie) que ceux formulés par l'utilisateur et que la requête représente ce besoin en information. Dans le cas du modèle booléen par exemple, le document sélectionné doit contenir tous les mots (cas conjonctif) ou une partie des mots (cas disjonctif) de la requête. Dans le cas du modèle vectoriel, plus

un document partage des mots avec la requête et dans la même proportion de poids, plus grande est sa similarité avec la requête. Ainsi, l'efficacité du procédé de sélection naïve de ces modèles, repose principalement sur l'efficacité et la qualité des mécanismes d'indexation et d'appariement [14]. Lors de l'appariement requête/document, seuls les documents qui sont les plus proches sémantiquement du besoin de l'utilisateur sont sélectionnés. De ce fait, plus les termes d'indexation sont représentatifs du contenu sémantique des documents et de la requête, plus la pertinence des documents sélectionnés est améliorée.

Néanmoins, dans la pratique la majorité des requêtes exprimées par les utilisateurs sont courtes et ambiguës [29], ce qui donne des spécifications inachevées sur leur besoin en informations. En outre, cette liste de termes ne correspond pas forcément à ceux utilisés pour indexer les documents pertinents de la collection et manque souvent de termes significatifs pouvant exprimer effectivement le besoin en information de l'utilisateur [10; 16; 50]. Ceci mène aux problèmes cruciaux de disparité des termes (*term mismatch problem*) [48] et d'ambiguïté [105] en recherche d'information (RI) : l'utilisateur et l'auteur d'un document n'utilisent pas nécessairement le même vocabulaire. Ainsi, un document peut être pertinent même s'il n'est ne contient pas les mêmes termes que ceux de la requête (plutôt des termes similaires). Cependant, dans les SRI classiques, un tel document ne sera aucunement retourné à l'utilisateur à cause du défaut d'appariement document-requête. De plus, une même requête exprimée par deux utilisateurs ayant des besoins différents va être exécutée de façon similaire par le SRI, et aucune distinction ne sera apportée aux résultats de recherche.

De ce fait, les performances d'un SRI, ne dépendent pas uniquement de l'efficacité et la qualité des mécanismes d'indexation et d'appariement, mais de façon non négligeable de la capacité du SRI de prendre en considération les besoins de l'utilisateur pour mieux répondre à leurs attentes. De ce constat est apparu un nouvel axe de recherche, celui de la RI *adaptative*.

1.4 La RI adaptative

Les travaux de la RI *adaptative* se sont particulièrement axés sur l'amélioration de l'efficacité du processus de recherche notamment lors de la phase d'exécution de la requête. Les techniques développées ont eu pour but de désambiguïser le sens des mots de la requête utilisateur afin de mieux cerner le but de sa recherche. Plus particulièrement, la RI *adaptative* s'articule autour de deux approches :

* Adaptation de la phase d'expression du besoin en reformulant la requête initiale de l'utilisateur.
* Adaptation du contenu informationnel du fond documentaire en identifiant des connexions appropriées entre les documents et les requêtes dans un domaine applicatif spécifique.

Dans ce contexte, on aborde dans ce qui suit les principales techniques de ces deux approches.

1.4.1 Reformulation de requête

La reformulation de requêtes est un processus ayant pour objectif de générer une nouvelle requête plus adéquate que celle initialement formulée par l'utilisateur en rajoutant de nouveaux termes et/ou supprimant des termes inutiles. Cette reformulation permet de coordonner le langage de recherche (utilisé par l'utilisateur dans sa requête) et le langage d'indexation des documents.

On distingue principalement deux approches de reformulation de requêtes : une approche basée sur un processus automatique et une autre, basée sur un processus interactif. Nous allons détailler dans les paragraphes suivants ces deux approches et nous présentons les principaux travaux développés.

1.4.1.1 Reformulation automatique de requête

La reformulation automatique de requête ou *expansion de requête* est l'une des premières techniques ayant produit des améliorations notables dans ce domaine. L'idée de base est d'ajouter à la requête initiale des termes issus de ressources linguistiques existantes ou bien de ressources construites à partir des collections.

Plus précisément, au niveau des ressources linguistiques, le but est d'utiliser un vocabulaire contrôlé issu de ressources externes. Il s'agit principalement de chercher des associations inter-termes extraites à partir des ontologies linguistiques (tel que WordNet [122]), ou à partir de thésaurus [188; 30].

Les thésaurus construits manuellement sont un moyen efficace pour l'expansion de requête. Cependant, leur construction et la maintenance des informations sémantiques qu'ils contiennent sont coûteuses en temps et nécessitent le recours à des experts des domaines considérés [188]. Ils sont de ce fait peu utilisés par les SRI. Les thésaurus construits automatiquement se basent essentiellement sur des méthodes de classification des termes [163; 139].

En ce qui concerne la seconde catégorie de ressources, elles sont construites en s'appuyant sur une analyse statistique des collections. Il s'agit de chercher des associations de termes afin d'ajouter des termes voisins à la requête. Il existe aussi d'autres méthodes entièrement automatiques telles que le calcul des liens contextuels entre termes [43] et la classification automatique de documents. Dans ce dernier cas, ces liens peuvent être construits à partir des documents retrouvés par le système, on parle alors de reformulation par contexte local, ou à partir de la collection entière de documents, on parle alors de reformulation par contexte global. Ces associations sont créées automatiquement et généralement basées sur la cooccurrence des termes dans les documents. Les liens inter-termes renforcent la notion de pertinence des documents par rapport aux requêtes

Dans le système INQUERY [87], la composante *PhraseFinder* crée automatiquement des associations entre les documents, les phrases sélectionnées sont directement utilisées pour l'expansion de requête. L'indexation sémantique latente (Latent Semantic Indexing) [52] a également été utilisée pour l'expansion de requête par analyse globale de contexte : se basant sur un espace dimensionnel de termes réduits pour représenter les documents de la collection, des corrélations entre ces termes peuvent être établies et utilisées pour augmenter la requête initiale.

L'analyse du contexte global exige donc de nombreuses statistiques sur de larges corpus de documents, des statistiques comme des mesures de cooccurrence entre les paires de termes, l'établissement de matrices de similarité entre les termes ou des associations dans des thésaurus de documents. Bien que cette approche soit relativement robuste, les mesures statistiques établies sur de larges corpus de documents et sur un très grand nombre de termes sont lourdes et coûteuses. Par ailleurs, cette approche ne prend en compte que les documents de l'ensemble de la collection qui ne sont pas forcément pertinents pour la requête. De ce fait, l'approche globale ne fournit seulement qu'une solution partielle aux problèmes d'ambiguïté des requêtes.

1.4.1.2 Reformulation interactive de requête

A la différence de la reformulation automatique, l'approche interactive (ou par réinjection de pertinence et/ou non-pertinence) exploite uniquement un sous-ensemble de documents sélectionnés parmi les premiers résultats obtenus de l'exécution de la requête initiale. Son principe fondamental est d'utiliser cette requête pour amorcer la recherche, puis modifier celle-ci à partir des jugements de pertinence et/ou de non pertinence de l'utilisateur, soit pour repondérer les termes de la requête initiale [144; 175], soit pour y ajouter (resp. supprimer) d'autre termes contenus dans les documents pertinents (resp. non pertinents) [146]. La nouvelle requête ainsi obtenue à chaque itération de *feedback*, permet de corriger la direction de la recherche dans le sens des documents pertinents.

Plusieurs techniques ont été introduites dans différents modèles de recherche [146; 106; 25; 23; 144; 171; 24] , notamment dans les modèles vectoriel et probabiliste, décrits ci-après.

1. **Réinjection de pertinence dans le modèle vectoriel**

   Dans le modèle vectoriel, la réinjection de pertinence consiste à rapprocher le vecteur requête à ceux des documents pertinents et l'éloigner des documents non pertinents. La nouvelle requête Q_{nlle} est construite grâce à la **formule de Rocchio** [146] dont l'idée est de dériver itérativement le vecteur requête optimal à partir d'opérations sur les vecteurs documents pertinents et les vecteurs documents non pertinents. Le vecteur de la nouvelle requête est construit comme suit :

$$Q_{nlle} = \alpha Q_{init} + \frac{\beta}{n_p}\sum_{n_p} D_p - \frac{\gamma}{n_{np}}\sum_{n_{np}} D_{np} \tag{1.3}$$

   Où : Q_{nlle} : vecteur de la nouvelle requête,

   Q_{init} : vecteur de la requête initiale,

   D_p (resp. D_{np}) : vecteur d'un document pertinent (resp. non pertinent),

   n_p (resp. n_{np}) : nombre de documents jugés pertinents (resp. non pertinents),

   α,β et γ : des paramètres constants, $\alpha+\beta+\gamma=1$

   Il est également possible de simuler l'interaction d'un utilisateur en postulant que les dix premiers documents trouvés par une première recherche sont pertinents et les suivants sont non pertinents (*pseudo relevance feedback*).

2. **Réinjection de pertinence dans le modèle probabiliste**

   Sur la base du modèle probabiliste, Robertson et Sparck-Jones [142] ont développé une formule de pondération des termes basée sur la distribution des termes de la requête dans les documents jugés pertinents et les documents jugés non pertinents par l'utilisateur. Cette formule est la suivante :

$$w_i = \frac{\frac{r_i}{R-r_i}}{\frac{n_i-r_i}{(N-n_i)-(R-r_i)}} \tag{1.4}$$

   Où :

   w_i : Poids du terme t_i dans la requête,

   r_i : Nombre de documents pertinents contenant le terme t_i,

   R : Nombre de documents pertinents pour la requête,

   n_i : Nombre de documents contenant le terme t_i,

   N : Nombre de documents dans la collection.

3. **Bilan de la réinjection de pertinence dans la reformulation de requête**

   Les premières expérimentations effectuées dans le cadre du système Smart [149] et poursuivies dans le modèle probabiliste [142] ont montré que la réinjection de pertinence permet l'amélioration de la précision des résultats essentiellement dans le cas de collections de petite/moyenne taille. Rocchio a également obtenu des résultats sensiblement positifs [146] lors de l'application de la réinjection de pertinence dans le modèle vectoriel.

   Par la suite, Salton et Buckley [150] ont effectué des expérimentations sur six collections de test pour comparer différentes méthodes. Ces méthodes sont principalement basées sur la repondération des termes de la requête. Typiquement, ces termes sont issus des documents jugés pertinents par l'utilisateur. Les résultats ont montré que les performances de recherche sont satisfaisantes dans la mesure où l'utilisateur fournit des jugements de pertinence de qualité et suffisamment précis durant plusieurs itérations de feedback. Ces conclusions suggèrent que l'efficacité du processus de reformulation de requête dépend fortement des dispositions des utilisateurs et leurs aptitudes à juger de la pertinence et/non pertinence des documents. Malheureusement, dans le contexte d'une recherche réelle les utilisateurs sont habituellement peu disposés à fournir un tel effort. En effet, l'exemple le plus parlant est celui du moteur de recherche *Excite* [58], qui avait mis en place ce type de reformulation de requête, mais qui n'a pas eu le succés excompté.

1.4.2 Adaptation du contenu documentaire

Dans cette approche, l'objectif de la RI adaptative est de définir des modèles de recherche, dits connexionnistes, permettant de décrire les représentations associatives entre les termes, les requêtes et les documents. L'idée de base est que les requêtes similaire ont un ensemble similaire de documents pertinents et que les informations capitalisées sur les documents pertinents pour ces requêtes devraient servir à retrouver les documents pertinents pour une nouvelle requête.

Les principaux travaux dans ce domaine se sont orientés vers l'application des réseaux de neurones. La particularité du réseau de neurones est de représenter les relations et associations qui existent entre les termes (ex. synonymie, voisinage, etc.), entre les documents (ex. similitude, référence, etc.), et enfin entre les termes et les documents (exemple, fréquence, poids, etc.).

Un réseau de neurone formel est construit à partir des représentations initiales des documents et de la requête. Le mécanisme de recherche d'information est fondé sur le principe de propagation de valeurs depuis les neurones descriptifs de la requête vers ceux des documents, à travers les connexions du réseau. Les résultats sont présentés à l'utilisateur selon le niveau d'activation des neurones documents. Le modèle connexionniste est connu pour sa capacité d'apprentissage, ce qui permet aux SRI de devenir adaptatifs.

Plusieurs modèles basés sur le principe des réseaux de neurones sont utilisés en RI [12; 106**?** ; 25; 47]. Cependant, il n'existe pas de représentation unique d'un réseau de neurones pour la recherche d'information, c'est au constructeur du modèle de le définir, et ce en identifiant les éléments suivants :

– Les différentes couches[1] du réseau (couche d'entrée, de sortie, intermédiaires, etc.) ;
– les neurones de chaque couche,
– la fonction d'entrée de chaque neurone,
– la fonction de sortie de chaque neurone,
– les liens entre les neurones et leurs poids associés.

Les travaux de [12] sont parmi les premiers a avoir abordé l'approche connexionniste en RI. AIR (Adaptive Information Retrieval), le système proposé dédié à la recherche dans le domaine bibliographique, est construit autour d'un réseau à trois couches : auteurs, termes et documents. Les liens entre les termes et les documents sont initialement pondérés par idf (le nombre de fois qu'un terme apparaît dans un document). Le système utilise les jugements des utilisateurs pour modifier ces liens dans le but d'arriver à une représentation consensuelle des termes dans les documents partagés par les utilisateurs.

Les modèles à couches, les plus performants de ces dernières années, sont ceux proposés par Kwok [106] dans le système de recherche d'information PIRCS et par Boughanem [21; 25] dans le système de recherche d'information MERCURE (ModèlE de Réseau Connexionniste poUr la REcherche d'information) :

Le modèle PIRCS est un réseau à couches interconnectées dans le sens requête(Q)-termes(T)-documents(D) [106]. Les connexions sont bidirectionnelles et asymétriques. L'approche de Kwok est fondée sur l'idée que les requêtes et documents sont similaires. Sur cette base, elle reprend des éléments du modèle probabiliste pour classer les neurones documents, répondant à une requête selon la probabilité, donnée par la formule suivante :

$$W_d = W_{qd} + W_{dq} \tag{1.5}$$

Où W_{qd} est la probabilité que la requête q soit pertinente pour le document d, et W_{dq} la probabilité pour que le document d soit pertinent pour la requête q.

Ces probabilités sont obtenues par propagation de signaux, dans le sens QTD pour obtenir W_{qd} et DTQ pour W_{qd}.

Dans le modèle connexionniste MERCURE, les requêtes, documents et termes sont représentés par des noeuds reliés entre eux par des liens pondérés.

[1] Une couche est un ensemble de neurones formels représentant un concept donné (requête, termes, documents, etc.)

Dans ce modèle trois types de liens pondérés sont proposés :

* Lien requête-termes pour modéliser le besoin en informations de l'utilisateur,
* lien terme-terme pour effectuer des reformulations directes de requêtes,
* lien terme-document.

L'appariement requête-document est effectué par un processus de propagation de signaux. Le processus est réalisé dans l'ordre suivant :

– La représentation de la requête est de la forme :

$$Q_u^{(t)} = (q_{u_1}^{(t)}, ..., q_{u_T}^{(t)})$$

Les poids des termes dans la requête sont affectés aux liens requête-termes.

– Déclenchement de l'évaluation à partir du noeud requête, en envoyant un signal de valeur 1 à travers les liens requête-termes.

– Calcul d'une valeur d'entrée et d'une valeur de sortie à chaque noeud :

$$In(t_i) = q_{ui}^{(t)}$$

$$Out(t_i) = g(In(t_i))$$

Où g est une fonction sigmoïde.

– Transmission des signaux vers la couche documents. Chaque noeud document calcule une entrée selon la formule :

$$In(d_j) = \sum_{i=1}^{T} Out(t_i) * dij$$

puis une valeur d'activation selon la formule :

$$Out(d_i) = g(In(d_i))$$

– Trier les documents répondant à la requête selon l'ordre décroissant de leur valeur d'activation.

MERCURE a été classé dans le top 5 des systèmes de recherche d'information ayant participés à l'édition TREC 97 [25; 23]. Le système permet également d'appliquer une reformulation de requête, selon deux modes distincts : une reformulation directe et indirecte. La reformulation directe consiste à ajouter de nouveaux termes à la requête initiale. Plus précisément, on ajoute les termes actifs, atteints par transfert d'activation à partir des termes de la requête. La reformulation indirecte se base sur l'injection de la pertinence dans le processus de recherche afin d'apprendre les liens requêtes-termes.

1.5 Bilan sur la RI adaptative : facteurs d'émergence de la RI personnalisée

Les travaux en RI adaptative ont certes apporté des solutions en particulier au défaut d'appariement requête-document qui ont conduit à l'amélioration des performances du processus de recherche d'information [71]. Cependant une analyse fine des travaux dans ce domaine montre que ces performances dépendent de nombreux facteurs *a priori* non contrôlés par le processus de réécriture adaptative de la requête. Ces facteurs, principalement liés aux approches de reformulation de requête, qui sont ainsi problématiques, peuvent être catégorisés selon trois principales dimensions : l'utilisateur, l'information portée par la requête et/ou document et l'interaction entre l'utilisateur et le SRI. Nous discutons dans ce qui suit chacune de ces dimensions [174].

1. *La dimension utilisateur*

   (a) l'expression initiale du besoin en information de l'utilisateur (ce qu'il ne sait pas du sujet de la requête) dépend de ses centres d'intérêt (ce qu'il sait déjà du sujet de la recherche) et de ses buts [79]. Cependant, ces éléments ne se déclinent pas dans le processus de réécriture de la requête initiale.

   (b) [76] montre qu'il existe une corrélation positive entre familiarité de l'utilisateur avec le sujet de la requête et les performances de la stratégie de réinjection de la pertinence. De plus, le niveau d'expertise de l'utilisateur [147; 193] a un impact sur les performances de recherche. En ce sens que des utilisateurs expérimentés effectuent de meilleurs choix quant à la qualité des documents et termes utilisés pour la réécriture de la requête, relativement à des utilisateurs novices.

   (c) [59] montre que la discipline professionnelle de l'utilisateur n'est pas sans impact dans la perception de l'information et donc de la pertinence. Ceci influe directement sur les performances de recherche.

   (d) la nature (utilité, intérêt, préférence) et valeur du jugement de pertinence de l'utilisateur (peu pertinent, très pertinent, assez pertinent etc.) dépend de nombreux facteurs : (1) de ses centres d'intérêt et ses buts [185] (2) de l'objet de la requête (ce qui est attendu à travers une requête : service, information, page de référence) [115; 89], (3) de la complexité de la tâche de recherche qui est déterminée par la quantité d'information que doit traiter l'utilisateur pour atteindre l'information pertinente [186]. Cependant, la RI adaptative exploite des jugements de pertinence binaire supposés ne dépendre que du contenu des documents.

2. *La dimension information*

   (a) le volume important d'information accessible engendre incontestablement une diversité importante du vocabulaire. Par conséquent, les algorithmes d'ordonnancement des termes d'expansion de requêtes en fonction de leur corrélation au sujet de la requête, sont peu performants [50].

   (b) les documents du Web contiennent de nombreuses informations non directement liées au sujet du document telles que les liens de navigation, les informations ou images publicitaires etc. Ces informations, même extraites des documents les mieux classés à l'issue d'une recherche initiale, engendrent du bruit lors d'un processus de réécriture de requête [199].

   (c) les stratégies classiques de réinjection de pertinence sont peu capables de rappeler des documents traitant de différents sujets auxiliaires associés à un sujet fédérateur véhiculé par la requête [203]. Le même problème est posé avec des documents traitant de nombreux sujets à la fois tels que les journaux [199].

3. *La dimension interaction*

   (a) les processus de réinjection de la pertinence induisent une interaction qui est à l'origine d'une surcharge cognitive pour l'utilisateur. La valeur ajoutée de ces interactions dépend du degré de participation de l'utilisateur. De plus, des études ont montré [13; 193] que les utilisateurs n'utilisent pas forcément l'ensemble des possibilités offertes par le système quant à l'enrichissement de la requête et ce, pour une raison majeure : les utilisateurs n'en cernent pas le principe et le lien avec l'opération de sélection de l'information pertinente.

   (b) la forme de présentation des documents (Titre, résumé, texte plein) exploités pour la réinjection de la pertinence a un impact non négligeable sur le jugement de l'utilisateur [81].

   (c) l'utilité de la réinjection de pertinence est plus déterminante aux dernières itérations d'un processus de recherche d'information adaptative [159].

Ce bilan montre globalement que les stratégies de RI adaptative ne sont pas garantes sur l'uniformité de la qualité des résultats d'un SRI dans des conditions d'utilisation différentes. Il en ressort que les éléments clés à intégrer dans de telles stratégies dans le but d'en améliorer les performances, sont dépendants les uns des autres, liés cependant à différentes dimensions. De ce fait, le développement de services d'accès délivrant l'information pertinente de manière personnelle en fonction des caractères spécifiques de l'utilisateur et adaptant les résultats de recherche en fonction des préférences et contexte de l'utilisateur devient une nécessité absolue. L'ensemble de ces éléments constitue l'ensemble des facteurs précurseurs ayant déterminé les directions d'investigation pour le développement de la troisième génération des SRI.

C'est pourquoi, au delà de la mise en œuvre des techniques d'adaptation, les travaux s'orientent actuellement vers la modélisation de l'utilisateur et son intégration comme composante du modèle global de recherche d'information. Ces travaux s'inscrivent dans le cadre précis de la « **personnalisation** de l'information ».

1.6 Conclusion

Au cours de ce chapitre, nous avons présenté les principaux facteurs d'émergence de la personnalisation dans le domaine de RI, à travers l'évolution des SRI classiques vers les SRIs adaptatifs. Nous avons présenté les principaux concepts de la RI, à travers l'architecture commune à tous les SRI, permettant l'appariement entre les requêtes formulées par des utilisateurs et les documents de la collection. Nous avons également présenté les différents modèles et stratégies utilisés lors de la mise en œuvre de ces concepts, caractérisant la première génération de SRI.

Mais, bien que ces stratégies reposent sur des modèles mathématiques, il n'en demeure pas moins que les résultats restitués par les SRIs classiques, compte tenu de l'important volume des ressources informationnelles disponibles, deviennent de moins en moins satisfaisants.

Ce constat a amené l'émergence d'une seconde génération de SRI, mené par le courant de la RI *adaptive*. Les premiers travaux de la RI *adaptative*, se sont particulièrement axés sur l'amélioration de l'efficacité du processus de recherche notamment lors de la phase d'exécution de la requête. Les techniques développées ont eu pour but de désambiguïser le sens des mots de la requête utilisateur et ajouter des termes pour mieux exprimer les besoins de l'utilisateur. Pour notre part, on s'est intéressé, dans ce chapitre, aux approches de reformulation de requête. Nous avons présenté les deux principales approches de reformulation de requêtes : La première est basée sur un processus automatique, la seconde est basée sur un processus interactif. L'objectif principal de ces approches est de générer une nouvelle requête plus adéquate que celle initialement formulée par l'utilisateur. Cette reformulation permet de coordonner le langage de recherche (utilisé par l'utilisateur dans sa requête) et le langage d'indexation des documents.

La reformulation de requête a effectivement permis d'améliorer les performances de la recherche. Cependant, du bilan présenté sur les stratégies de la RI adaptative on constate que, vu le contexte actuel lié au volume d'information, ces techniques sont peu viables. En effet, pour que la reformulation soit performante, elle nécessite une forte implication de l'utilisateur ce qui peut engendrer une surcharge cognitive importante. De plus, la stratégie de recherche n'est pas adaptée à un contexte d'utilisation dans différents domaines d'intérêts ; le contexte de l'utilisateur demeure peu connu et par conséquent de moindre impact sur le processus de recherche. De plus, un même utilisateur peut avoir différents besoins à différents instants.

Force est de constater qu'il n'y a généralement pas de mécanisme explicite qui représente et intègre l'utilisateur dans le processus de recherche, l'utilisateur est entièrement représenté par sa requête. Cette problématique est d'autant plus accentuée que la majorité des requêtes exprimées par les utilisateurs sont courtes et ambiguës [29], ce qui donne des spécifications inachevées sur leur besoin en information. Lorsque l'utilisateur initie la recherche, c'est dans le but de combler un manque en information sur un sujet précis. Ce besoin en information est fortement lié à différents facteurs dépendants, d'une part, du but de sa recherche (son activité, son contexte et son environnement) et d'autre part, des connaissances capitalisées par l'utilisateur (ses préférences et centres d'intérêts) tout au long de ses expériences. Ces différents aspects liés à l'utilisateur s'inscrivent dans le cadre précis de la « personnalisation de l'accès à l'information » [119] que nous abordons dans le chapitre suivant.

2 L'accès personnalisé à l'information : Préambule et Problématique

2.1 Introduction

La personnalisation de l'accès à l'information peut être définie comme un processus qui *change la fonctionnalité, l'interface, la teneur en information, ou l'aspect d'un système pour augmenter sa pertinence personnelle en fonction des caractéristiques socio-démographiques déclarées de l'utilisateur (sexe, âge, lieu de résidence, etc.) et/ou de son comportement observé* [182].

Ce chapitre introduit la thématique de recherche de la personnalisation de l'information à travers les éléments clés de la personnalisation liés à l'entité *utilisateur*. La section 2.2 présente l'objectif général de la personnalisation. La section 2.3 traite de la formalisation du contexte en RI. Nous présentons en premier différentes définitions du contexte ainsi que ses principaux éléments retenus dans la littérature. Nous présentons ensuite à travers l'architecture générale d'un système d'accès personnalisé à l'information, les différentes phases du cycle de vie d'une requête où intervient le profil de l'utilisateur. La section 2.4 aborde les principales questions posées par la communauté en RI qui constituent les verrous scientifiques et technologiques de la personnalisation. La section 7.4 conclura ce chapitre.

2.2 Préambule

La démocratisation des moyens informatiques dans tous les secteurs d'activité humaine et notamment comme outil de communication, font émerger la personnalisation comme approche essentielle aux succès des systèmes d'accès à l'information.

En effet, face aux phénomènes actuels d'accroissement incessant d'informations ainsi qu'à leur hétérogénéité, s'impose de nouvelles réflexions sur les méthodologies de conception et de développement de la troisième génération des systèmes d'accès à l'information[1].

[1]Après la RI classique, puis la RI adaptative, considérée comme la permière et seconde génération de SRI

Dans ce contexte, [198] décrivent trois axes de réflexion. D'abord, le système doit avoir la capacité de détecter l'intention de recherche de l'utilisateur ; deuxièmement, il doit offrir à l'utilisateur des capacités et des services améliorés pour fournir plus d'informations lors de l'expression de ses besoins, qu'une simple requête ; et troisièmement, il doit pouvoir mettre en œuvre des interactions et des mécanismes plus sophistiqués avec l'utilisateur pour réaliser les deux premiers points.

La personnalisation est considérée dès lors comme un aspect dominant dans plusieurs secteurs des multimédia interactifs. Elle peut cibler les deux aspects d'un système (interface ou contenu) de manière spécifique ou simultanément :

* Au niveau de la présentation : personnaliser des aspects de l'interface utilisateur, y compris les couleurs, les polices, le positionnement et l'affichage des données. Cet aspect correspond à ce que l'on nomme la *customisation* des systèmes. Les aspects de mobilité et d'environnement géographique des utilisateurs sont également pris en considération.
* Au niveau du contenu : cibler la recherche en fonction des besoins et des centres d'intérêts des différents utilisateurs. Cet aspect correspond plus à ce que l'on nomme « *personnalisation* ».

Dans le cadre d'un système dédié exclusivement à l'accès à l'information, l'objectif de la personnalisation est d'intégrer l'utilisateur dans tout le processus de recherche afin de lui délivrer une information pertinente en fonction de ses caractéristiques spécifiques.

A ce stade, une distinction entre ces deux aspects est nécessaire. Dans le cas de la customisation, le système ajuste son interface, sa structure et représentation aux préférences de chaque utilisateur. A chaque connexion d'un utilisateur, le système doit charger l'interface adaptée à ses besoins. Dans le cas de la personnalisation du contenu, le système vise à augmenter le processus de recherche initié explicitement par les requêtes de l'utilisateur avec des caractéristiques informationnelles extraites explicitement/implicitement de l'utilisateur, dans le but d'améliorer ses différents besoins [65].

Néanmoins, indépendamment de ces deux types de personnalisation, les principaux éléments communs à tous les systèmes d'accès incluent :

(a) la catégorisation et le prétraitement des données de recherche et/ou de navigation,

(b) l'extraction des corrélations et liens existants entre ces différents types de données, et

(c) la détermination des actions qui devraient être effectuées par un tel système de personnalisation [127].

Tous ces critères suggèrent l'importance du contexte de recherche et son intégration dans le processus d'accès à l'information. Plusieurs études ont effectivement montré que la principale raison de l'insatisfaction des utilisateurs demeure l'aspect non personnalisable du processus d'accès à l'information. Dans la majorité des systèmes, le contexte de recherche de l'utilisateur est peu connu, voir uniquement représenté par la requête de l'utilisateur.

(respectivement)

Ainsi, toute information sur l'utilisateur, comme ses préférences, ses centres d'intérêts, ses besoins en information et son environnement de recherche sont de ce fait supposés pertinents et exploitables par le système de personnalisation. L'ensemble de ces informations va correspondre à ce que l'on nomme le **contexte de l'utilisateur** ou dans un cadre plus spécifique **profil utilisateur**.

Ces deux notions sont introduites dans la section suivante. Il est à noter que la notion de contexte est générale et englobe plusieurs dimensions informationnelles, pour notre part on s'intéresse à la notion de profil utilisateur, qui correspond à une des dimensions du contexte.

2.3 Notions de base

2.3.1 Contexte de recherche

Il a été largement admis que la troisième génération des systèmes d'accès à l'information doit prendre en considération le contexte de recherche des utilisateurs [2; 46; 60] dans tout le processus d'accès à l'information.

Crestani et Ruthven [49] stipulent que : « *The context affects all aspects of information retrieval. A searcher's context affects how they interact with a retrieval system, what type of response they expect from a system and how they make decisions about the information objects they retrieve.*»

Dans ce cadre, le terme *contexte* se décline selon plusieurs facteurs. On ne trouve pas dans la littérature de définition complète et générique de la notion du contexte et plus précisément des éléments qui le constituent. Les facteurs tels que la date, le lieu, l'historique d'interaction, la tâche en cours, l'environnement de recherche et d'autres facteurs implicites non explicités dans l'interaction, ont été largement exploités dans diverses domaines de recherche, à savoir, l'intelligence artificielle, la RI, les bases de données, l'image et l'analyse de vidéo, etc.

Les premiers travaux de Saracevic [155] et Ingerwersen [79] définissent le contexte selon un modèle cognitif par lequel on peut identifier des structures ou espaces cognitifs qui sont autant de variables impliquées dans le processus de RI et qui peuvent décrire les intentions et les perceptions de l'utilisateur et de ce qui l'entoure. Ces variables sont l'espace cognitif de l'utilisateur, l'environnement social ou organisationnel, les intentions et les buts de l'utilisateur ainsi que le système lui-même.

Par la suite, [46] définissent des niveaux contextuels considérés les plus significatifs en RI afin de dissocier les entités qui interviennent dans le processus de recherche :

* le premier niveau concerne l'environnement de recherche lié aux facteurs cognitifs, sociaux ou professionnels qui influent sur le comportement de recherche de l'utilisateur et sa perception de la pertinence,
* le deuxième niveau concerne la RI liée aux connaissances de l'utilisateur : ses buts et ses intentions de recherche,
* le troisième niveau concerne l'interaction utilisateur-système et met en relief l'impact des situations ou de l'environnement sur la rétro-action ou les jugements de pertinence de l'utilisateur,
* le dernier niveau concerne le niveau de requête ou le niveau linguistique du contexte ; ce niveau explore la performance du SRI dans l'interprétation des requêtes des utilisateurs et leur habilité à les désambiguïser.

Un contexte multidimensionnel a également été défini par [64]. Cette définition ajoute de nouvelles caractéristiques liées d'une part à l'aspect temporel du besoin en information et d'autre part au type de recherche demandé. Les trois principales dimensions retenues pour le contexte sont : social, application et temps. La dimension sociale définit l'appartenance possible de l'utilisateur : individuel, groupe ou communauté. La dimension application définit le but de la tâche accomplie : recherche ad-hoc, résolution du problème. La dimension temps permet de définir le contexte temporel du besoin : temps passé (*batch*), intention à court terme ou intention à long terme. Le contexte à court terme (*interactif*) ou courant est associé aux besoins et préférences de l'utilisateur lors d'une session de recherche, alors que le contexte à long terme (*personnalisation*) traduit les besoins et les préférences persistants de l'utilisateur tout au long de diverses sessions de recherche.

Dourish [56] présente également deux aspects de contexte. Le premier considère le contexte comme étant une «*forme d'information*» qui se caractérise par sa «*stabilité*». Le contexte est vu comme un ensemble d'éléments environnementaux liés aux activités génériques de l'utilisateur. Les travaux adoptant cette perspective, ont été principalement confrontés aux difficultés liées à la manière de capturer et de représenter un contexte stable. Le deuxième aspect considère le contexte comme « *un dispositif émergeant de l'interaction, déterminé par le temps et le contenu* ». En effet, Dourish déclare que «*context and content (or activity) cannot be separated. Context cannot be a stable, external description of the setting in which activity arises. Instead, it arises from and is sustained by the activity itself*».

Plus récemment, Ingwersen et Järvelin [80] proposent un modèle basé sur des infrastructures incluant de larges classes contextuelles. L'objectif de cette infrastructure de recherche cognitive est d'étudier les dimensions du contexte qui ont un impact sur le processus de RI. Elle se compose de 9 classes/dimensions du contexte et dépendantes les unes des autres.

On peut citer :

- la dimension de la tâche naturelle de travail dans une organisation ou une collaboration concerne les caractéristiques d'intérêts liées à la tâche,
- la dimension de la tâche de recherche liée à la tâche naturelle de travail,
- la dimension de l'utilisateur qui concerne ses caractéristiques et ses connaissances déclarées, sa perception quant à la tâche de travail et la tâche de recherche y inclut les différents types de besoins en information en fonction de la tâche de travail,
- les caractéristiques de la collection de documents (genre des documents, etc.),
- les caractéristiques du système tel que le principe de représentation des documents et des besoins et les méthodes utilisées pour leur appariement.

Ils proposent également une classe historique représentant l'évolution du contexte lors de la recherche ; cette dimension traduit les changements survenus dans le contexte de l'utilisateur à chaque nouvelle recherche, lorsque de nouveaux résultats sont examinés ou qu'un nouveau document est sélectionné.

Il en ressort qu'indépendamment des différentes définitions données au contexte, on trouve un consensus commun regroupant des dimensions descriptives tel que l'environnement cognitif, le besoin en information, les centres d'intérêts récurrents et l'interaction liée à l'activité d'accès.

2.3.2 Profil utilisateur

La notion de profil utilisateur a été largement abordée dans le domaine du *user modeling*. Depuis le début des années 70, les recherches menées dans ce domaine se sont principalement focalisées sur la possibilité de définir des approches de modélisation de l'utilisateur dans le contexte de différentes applications [137]. L'objectif de ces approches est d'améliorer les interactions homme-machine (IHM) par inférence et prédiction des buts, préférences et contextes des utilisateurs à partir de faits observés.

Le concept de profil utilisateur a été introduit pour l'accès à l'information en premier dans les travaux de filtrage d'information [14], pour décrire une structure représentative de l'utilisateur, en l'occurrence ses centres d'intérêts. Cette notion a ensuite été réexploitée en RI personnalisée pour former les composantes du contexte directement dépendantes de l'utilisateur : centres d'intérêts, préférences, domaines professionnels, expertise, etc.

On appelle **profil utilisateur toute structure qui permet de modéliser et de stocker les données caractérisant l'utilisateur**. Ces données représentent les centres d'intérêts, les préférences et les besoins en informations de l'utilisateur ou un groupe d'utilisateurs [200], [28].

Il convient de distinguer la notion de profil de la notion de requête. Un profil peut être défini comme une mise en équation du centre d'intérêt et des préférences de l'utilisateur, alors qu'une requête est l'expression d'un besoin circonstancié que l'utilisateur souhaite voir satisfait en tenant compte de son profil. Un profil a un caractère plus invariant que les requêtes même si le centre d'intérêt et les préférences de l'utilisateur peuvent légitimement évoluer. [99]

2.3.3 Pertinence contextuelle

La pertinence est incontestablement la question fondamentale posée lors de l'accès à l'information. La pertinence est une notion subjective, non généralisable à tout type d'information et valide pour tout type d'utilisateurs. Elle dépend notamment du centre d'intérêt ou domaine d'application, du moment, du lieu et du support que l'utilisateur a choisi pour accéder à l'information et du système qui délivre cette information.

Cette notion subjective, dépendant essentiellement du point de vue de l'utilisateur, a de nouveau été l'objet d'investigations dans le cadre de l'accès personnalisé à l'information de manière générale et de la RI personnalisée de manière particulière. Dans ce contexte, la pertinence est spécifiée comme étant un concept multidimensionnel [18], dont on distingue principalement quatre types :

- *pertinence algorithmique* : la pertinence est traduite par une mesure algorithmique dépendant des caractéristiques des requêtes d'une part et des documents d'autre part. C'est le seul type de pertinence qui est indépendant du contexte ;
- *pertinence thématique* : la pertinence traduit le degré d'adéquation de l'information à couvrir, en partie, le thème évoqué par le sujet de la requête. C'est le type de pertinence adressé par les assesseurs de la campagne d'évaluation TREC ;
- *pertinence cognitive* : c'est la pertinence liée au thème de la requête, « pondérée » par la perception ou les connaissances de l'utilisateur sur ce même thème ;
- *pertinence situationnelle* : c'est la pertinence liée à la tâche de recherche. Ce type de pertinence traduit essentiellement l'utilité de l'information relativement au but de recherche de l'utilisateur.

La RI personnalisée explore essentiellement la pertinence cognitive et la pertinence situationnelle.

2.3.4 Architecture fonctionnelle

Un système de recherche d'information personnalisé (*SRIP*) est un système qui intègre l'utilisateur, en tant que structure informationnelle, tout au long de la chaîne d'accès à l'information. Le *SRIP* ne se limite pas seulement à modéliser les caractéristiques des utilisateurs en des profils. Il doit être capable de déduire à partir de ces profils, l'intention de l'utilisateur lorsqu'il effectue sa recherche, en d'autres termes son contexte de recherche, et de détecter

l'évolution des profils de manière dynamique. Le système doit donc inclure :

* des techniques et algorithmes pour capturer et modéliser le but, les préférences et les centres d'intérêts de l'utilisateur ou un groupe d'utilisateurs. Un modèle de profil utilisateur est alors décrit et instancié,
* une procédure de mise à jour du profil qui traduit son évolution dans le temps,
* des mécanismes et algorithmes pour intégrer le profil de l'utilisateur dans le processus d'accès et retourner l'information pertinente en fonction de ce profil.

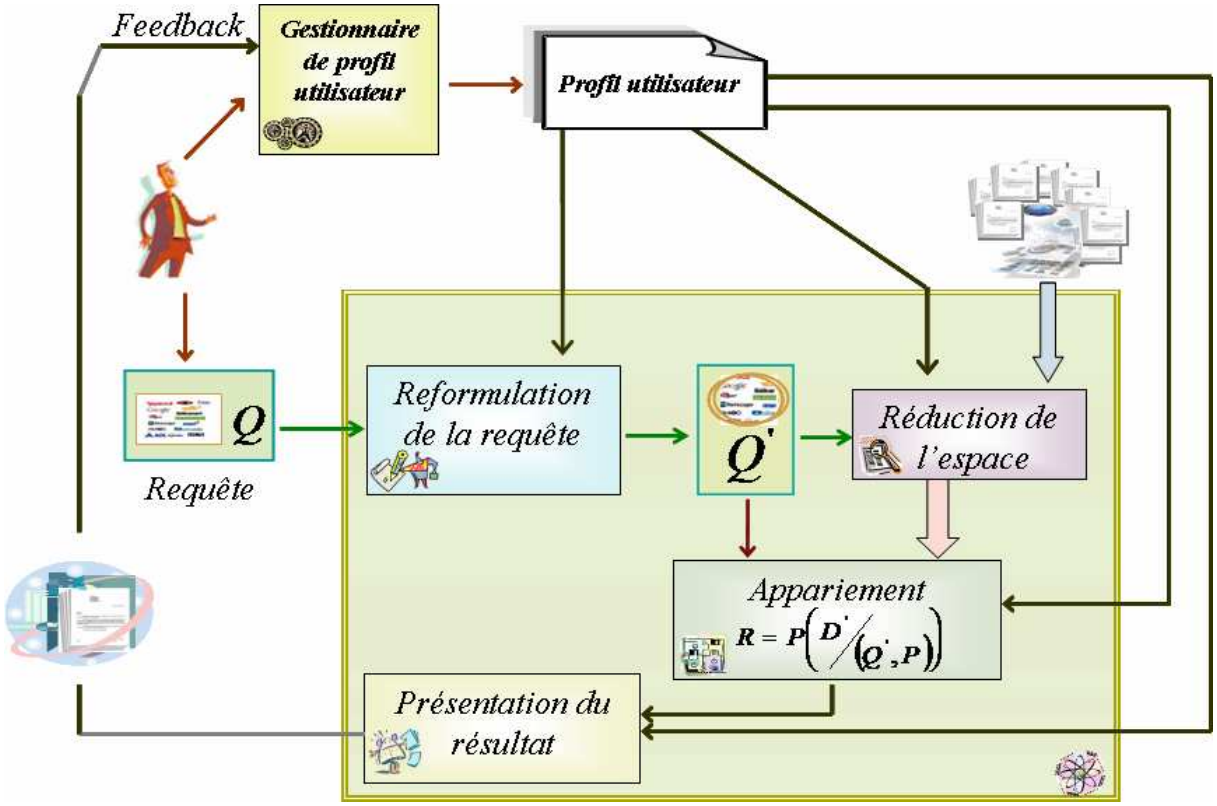


Figure 2.1 — Processus global d'accès personnalisé à l'information

La diversité des systèmes de personnalisation rend la définition d'une architecture fonctionnelle et formelle d'accès personnalisé à l'information difficilement généralisable. Néanmoins, nous tentons dans ce paragraphe de dégager une architecture standard pour un système d'accès personnalisé à l'information, présentée par la figure 2.1. Notre volonté est de mettre en évidence l'ensemble des fonctionnalités de tels systèmes même si elles ne sont pas toujours présentes collectivement dans tout système.

Cette architecture centrée autour de l'utilisateur met en évidence :

1. *Un gestionnaire de profil* pour représenter, construire et faire évoluer les profils des utilisateurs.
2. Les étapes du cycle de vie de la requête où l'on intègre le profil utilisateur dans :
   (a) *la phase de reformulation de la requête* afin de mieux cibler le contexte de la recherche de l'utilisateur,
   (b) *la phase de réduction de l'espace de recherche* pour restreindre l'espace de recherche aux documents qui ciblent les besoins de l'utilisateur,
   (c) *la phase d'appariement* pour calculer la pertinence des documents en fonction des caractéristiques spécifiques de l'utilisateur,
   (d) *la phase de présentation des résultats* pour restituer les informations selon le contexte et les préférences de l'utilisateur.

2.4 Problématique générale

En tant que thématique de recherche, la personnalisation de l'accès à l'information est confrontée à des verrous que l'on peut projeter sur deux niveaux. Le premier niveau est lié à la modélisation du profil de l'utilisateur. Cette étape consiste à définir la structure représentative des informations contenues dans le profil d'une part et d'instancier cette structure (construction et évolution) pour chaque utilisateur, d'autre part. Le second niveau se rapporte à l'exploitation de la dimension utilisateur lors de la mise en œuvre du processus d'accès personnalisé à l'information.

2.4.1 Modélisation du profil utilisateur

L'introduction de la dimension utilisateur dans un processus d'accès à l'information, mérite, voire nécessite une réflexion sur la modélisation de l'entité *utilisateur*. La fiabilité ou qualité des profils est en effet d'une importance bien connue dans le domaine de la modélisation utilisateur (*User modeling*) [98]. En effet, on constate que l'une des principales raisons du manque de performances des techniques de personnalisation est typiquement l'application d'un profil utilisateur hors contexte [66].

Les utilisateurs peuvent avoir des préférences générales, récurrentes et stables. Cependant, l'ensemble des informations contenues dans le profil ne sont pas forcément appropriées à toutes les situations de recherche. Le plus souvent, les systèmes n'utilisent seulement qu'un sous-ensemble de ces informations, qu'ils supposent pertinents pour la recherche en cours. Dés lors, le choix du profil adéquat constitue la principale réflexion lors de la mise en œuvre du SRIP.

De ce fait, les questions fondamentales posées pour modéliser le profil utilisateur sont le « *Quoi* », le « *Comment* » et le « *Quand* » [4] :

* Quoi ?
  - Quelles propriétés informationnelles caractérisent l'utilisateur ?
  - Quel structure informationnelle utiliser pour représenter l'utilisateur ?
* Comment ?
  - Comment collecter les informations du profil ?
  - Comment construire le profil de l'utilisateur ?
  - Comment détecter le contexte, le but de la recherche et les besoins à court/long terme de l'utilisateur ?
  - Comment adapter le profil à l'évolution de l'utilisateur lui même ?
  - Comment assurer la sécurité et la confidentialité des informations du profil ?
* Quand ?
  - Quand faut-il faire évoluer le profil de l'utilisateur ?

2.4.2 Exploitation du profil utilisateur

Le domaine d'application conditionne fortement les techniques de personnalisation employées par le système. Cela a en effet un impact direct sur l'exploitation du profil dans la chaîne d'accès à l'information et par conséquent sur les mécanismes mis en œuvre. Les principales interrogations posées concerne le « Quoi » , le « Comment » de la mise en œuvre :

* Quoi ?
  - Quels services de personnalisation proposer : de la recommandation et/ou du filtrage, de l'aide à la navigation, un assistant personnel de recherche ?
  - Dans quelles étapes du cycle de vie de la requête faut-il intégrer le profil ?
  - Quelles informations du profil exploiter lors de l'accès à l'information ?
* Comment ?
  - Comment intégrer le profil de l'utilisateur dans le processus de personnalisation ?
  - Comment évaluer l'impact de la personnalisation sur le processus de recherche ?

2.5 Conclusion

Ce chapitre a permis de cerner la thématique de la personnalisation de l'accès à l'information dans sa globalité. On y présente en premier les éléments de base autour de la notion de l'utilisateur : contexte de recherche, profil utilisateur, pertinence contextuelle. Puis, à travers l'architecture générale du processus d'accès, nous avons identifié les principales fonctionnalités de la personnalisation : la gestion du profil, et les phases du cycle de vie de la requête, où l'on intègre l'utilisateur.

Le chapitre aborde également la problématique générale de la personnalisation. Elle s'articule autour de ces principales questions : *Quel structure choisir pour représenter le profil ? Quel est son contenu informationnel ? Comment collecter ces informations et enfin comment le système les exploite dans le processus d'accès à l'information*. Les approches de personnalisation proposées dans la littérature tendent à répondre à ces questions. Dans la suite de cette thèse, nous abordons dans le chapitre 3 les principales approches de modélisation de l'utilisateur. Les différents modèles d'accès personnalisé à l'information seront présentés dans le chapitre 4,

3 L'accès personnalisé à l'information : Modélisation de l'utilisateur

3.1 Introduction

La modélisation de l'utilisateur est au cœur de la mise en œuvre de processus d'accès personnalisé à l'information. Elle consiste à décrire les caractéristiques informationnelles des utilisateurs à travers un modèle de profil. On trouve dans la littérature plusieurs définitions de la modélisation utilisateur, dont nous citons quelques unes ci-dessous :

"*A user model is a knowledge source in a natural-language dialogue system which contains explicit assumptions on all aspects of the user that may be relevant for the dialogue behavior of the system.* " [100]

"*The process of gathering information about the users of computer systems and of making this information available to systems which exploit it to adapt their behavior or the information they provide to the specific requirements of individual users has been termed as user modeling.*" [137]

"*User model is an explicit representation of the system of a particular user's characteristics that may be relevant for personalized interaction.*" [141]

Pour notre part, nous considérons que la modélisation de l'utilisateur ; dans le cadre d'un système d'accès personnalisé à l'information, est un processus caractérisé par trois phases :

– la première porte sur la définition d'une représentation des unités d'information caractérisant l'utilisateur du système. Elle correspond à la définition du profil utilisateur,
– la deuxième phase est liée à l'instanciation de cette représentation au cours d'une activité d'accès à l'information pour un utilisateur particulier. Elle regroupe des techniques d'acquisition des données utilisateur ainsi que des approches de construction pour agencer les informations collectées selon la structure représentative définie lors de l'étape précédente,

– enfin, la troisième phase concerne l'évolution du profil au cours du temps. Elle nécessite la mise en place de stratégies de mise à jour du contenu informationnel du profil.

Plusieurs techniques furent développées dans la littérature pour modéliser l'utilisateur, elles diffèrent cependant dans l'approche de représentation et construction du profil. En effet, le contenu informationnel du profil dépend fortement de l'application. Le plus souvent ce sont les données exploitées par le système qui déterminent le contenu du profil. La plupart des travaux actuels en RI se focalisent à juste titre, sur la représentation de l'aspect lié aux intentions de l'utilisateur qualifiées de centres d'intérêts. Dans cette perspective, la modélisation du profil de l'utilisateur a pour objectif fondamental de représenter puis faire évoluer ses besoins en information à court et moyen terme. C'est une question qui pose la double difficulté de traduire les centres d'intérêt de l'utilisateur d'une part et faire émerger leur diversité d'autre part. En outre, l'acquisition du profil de l'utilisateur dépend des mécanismes utilisés par le système pour formuler des prédictions au sujet de l'utilisateur et inférer de son comportement les informations traduisant ses préférences et ses centres d'intérêts.

Ce chapitre est consacré à la modélisation de l'utilisateur pour l'accès personnalisé à l'information. Nous y présentons les principales phases du processus de modélisation : les approches de représentation dans la section 3.2, les approches de construction dans la section 3.3 et les méthodes d'évolution des profils utilisateurs dans la section 3.4. Enfin une conclusion générale sera présentée dans la dernière section.

3.2 Approches de représentation du profil utilisateur

La représentation de l'utilisateur à travers la notion de profil permet de mieux comprendre certains mécanismes cognitifs, notamment ceux permettant de percevoir le concept subjectif de la pertinence et au-delà, cibler ses besoins spécifiques dans le but d'améliorer les performances de recherche. Le profil de l'utilisateur, constitué de paquets divers d'informations le caractérisant, traduit une connaissance éparse sur l'utilisateur. Dans le cadre de la RI, l'unité élémentaire utilisée pour représenter ces paquets d'informations est le **terme pondéré**. Un modèle de représentation permet d'organiser ces éléments afin de faciliter leur exploitation dans le processus d'accès à l'information. On distingue quatre principales approches de représentation : ensembliste, connexionniste, conceptuelle et multidimensionnelle.

3.2.1 Représentation ensembliste

L'approche ensembliste consiste à représenter le profil de l'utilisateur par des paquets de termes pondérés. D'un point de vue RI, on parle plutôt d'une représentation vectorielle par analogie au modèle vectoriel de Salton [149] sur laquelle elle se base. Ces paquets de termes, traduisant les centres d'intérêts de l'utilisateur, peuvent être regroupés différemment selon l'approche suivie pour considérer le profil de l'utilisateur.

On distingue dans la littérature trois grandes approches de représentation du profil utilisateur basées sur ce modèle :

* Par une liste de mots clés, où chaque mot correspond à un centre d'intérêt spécifique [5].
* Par un vecteur de termes pondérés pour chaque centre d'intérêt [39; 179].
* Par un ensemble de vecteurs de termes pondérés (ou non) indépendants, pour prendre en compte des centres d'intérêt multiples [162] où chaque vecteur correspond à un domaine d'intérêt [134] .

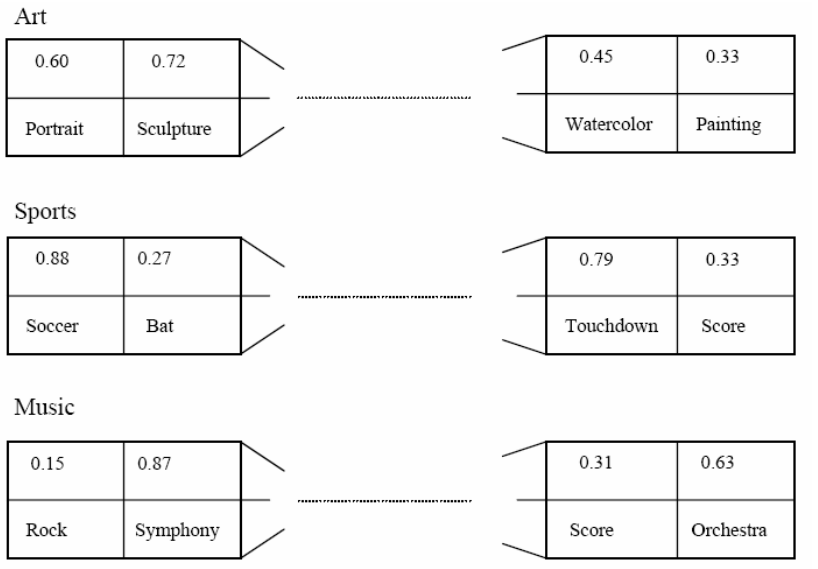


Figure 3.1 — Un exemple de profil représenté par des mots clés

La représentation ensembliste fut parmi les premiers modèles de profils utilisateur exploités en RI. La pondération des termes est généralement basée sur un schéma de la forme $TF*IDF$ communément utilisé en RI [153]. Le poids associé à chaque terme permet de représenter son degré d'importance dans le profil de l'utilisateur. La figure 3.1 donne un exemple de profil utilisateur représenté par des mots clés pondérés. Ce profil contient trois centres d'intérêts : *Art*, *Sports* et *Music*. Chaque centre est représenté par un ensemble de termes pondérés. $Music =< (Rock,\ 0,15),\ (Symphony,\ 0,87), \cdots >$ est un extrait du l'ensemble de termes pondérés représentant le centre *Music*.

Plusieurs systèmes d'accès personnalisé à l'information utilisent ce type de représentation. Notamment, dans Anatagonomy [148], un système personnalisé de consultation de nouvelles et de journaux en ligne, Fab [8] un système de recommandation de page *web*, Letizia [110], un système d'aide à la navigation, et Syskill & Webert [134] un système de recommandation. Tous ces systèmes proposent des profils utilisateur représentés par une liste de mots clés.

Dans le même cadre, on trouve PEA [128], un système d'aide personnalisée à la navigation qui établit des profils utilisateur basés sur la représentation vectorielle, en utilisant des termes extraits des pages annotées par l'utilisateur lors de sa navigation [34]. Cependant,

à la différence des autres systèmes, plutôt que de créer un profil unique pour l'utilisateur, dans PEA, l'utilisateur est représenté par un ensemble de vecteurs de termes pondérés, pour chaque annotation. Leur principe de base, est que l'utilisateur peut avoir plusieurs centres d'intérêts lors de sa recherche. La combinaison des termes représentant ces centres dans un même vecteur permet d'obtenir un profil couvrant l'ensemble de ses centres.

WebMate [39] établit également des profils d'utilisateur contenant un vecteur de terme par centre d'intérêt, tandis qu'Alipes [194] augmente cette approche en représentant chaque intérêt avec trois vecteurs de termes : un descripteur à long terme et deux descripteurs à court terme : un négatif et un second positif (représentant les centres non intéressants et intéressants de l'utilisateur, respectivement).

La représentation par liste de mots clés et/ou par classe de vecteurs de termes apporte l'avantage de la simplicité de mise en oeuvre. Néanmoins, même si ces systèmes prennent en considération des centres d'intérêts multiples en utilisant plusieurs vecteurs, cette représentation manque de structuration. Elle ne facilite ni l'interprétation ni la prise en compte des différents niveaux de généralité caractérisant l'utilisateur [20]. En effet, la plupart des utilisateurs ont des intérêts divers et multiples, leur généralisation dans un vecteur simple n'est pas clairement représentative de la réalité.

En outre, une représentation par des vecteurs multiples regroupant les documents intéressants l'utilisateur sur la base de leurs similitudes, est sensée être représentative du même centre d'intérêt. Néanmoins, l'efficacité des profils dans cette approche dépend fortement du degré de généralisation pour représenter de tels intérêts. Le problème est lié à l'application d'une analyse statistique des mots-clés indépendamment de toute information contextuelle. En d'autres termes, les documents sont considérés comme indépendants sans tenir compte de la situation contextuelle lors de la recherche de l'utilisateur. Il reste aussi à résoudre le problème de l'ordonnancement des préférences et des centres d'intérêts de l'utilisateur. En effet, ces derniers sont très variés et n'ont pas le même degré d'importance pour chaque utilisateur. Dans le cas d'une représentation en hiérarchie de classes ou de concepts, le rapport de généralisation/spécification existant naturellement dans ce genre de structure permet d'avoir une représentation plus réaliste du profil utilisateur.

3.2.2 Représentation connexionniste

Afin d'adresser le problème de polysémie des termes inhérents à la représentation ensembliste, une première solution consiste à représenter le profil par un réseau de nœuds pondérés dans lequel chaque nœud représente un concept traduisant un centre d'intérêt utilisateur. Ce type de représentation offre le double avantage de la structuration et de la représentation associative permettant de considérer l'ensemble des aspects représentatifs du profil.

Les centres d'intérêts sont souvent représentés par des relations de paires de nœuds dans laquelle chaque nœud contient un terme issu des documents du corpus de recherche. Les arcs reliant deux nœuds, sont créés sur la base des co-occurrences entre ces termes. Le pro-

fil utilisateur peut être augmenté par l'inclusion d'un ensemble de paires d'attribut-valeur correspondant à la partie structurée des documents [123] (par exemple, la taille, nombre d'images, etc.) qui ont précédemment intéressé l'utilisateur [8].

Cependant, la représentation séparée de chaque mot par des nœuds dans le réseau sémantique n'était pas assez précise pour décliner les différentes significations des centres d'intérêts de l'utilisateur. Une alternative possible est d'exploiter des sources externes telles que les ontologies pour établir les liens entres les nœuds.

Dans ce cadre, le système SiteIF [168] propose d'utiliser les concepts inhérents à WordNet pour regrouper des termes semblables dans des concepts appelés des "*ensembles de synonyme*", ou des *synsets*. Le profil de l'utilisateur est alors représenté comme un réseau sémantique dans lequel les nœuds sont des synsets, les arcs sont des co-occurrences des membres de synset avec le document intéressant l'utilisateur. Les nœuds et les poids des arcs représentent le niveau d'intérêt de l'utilisateur.

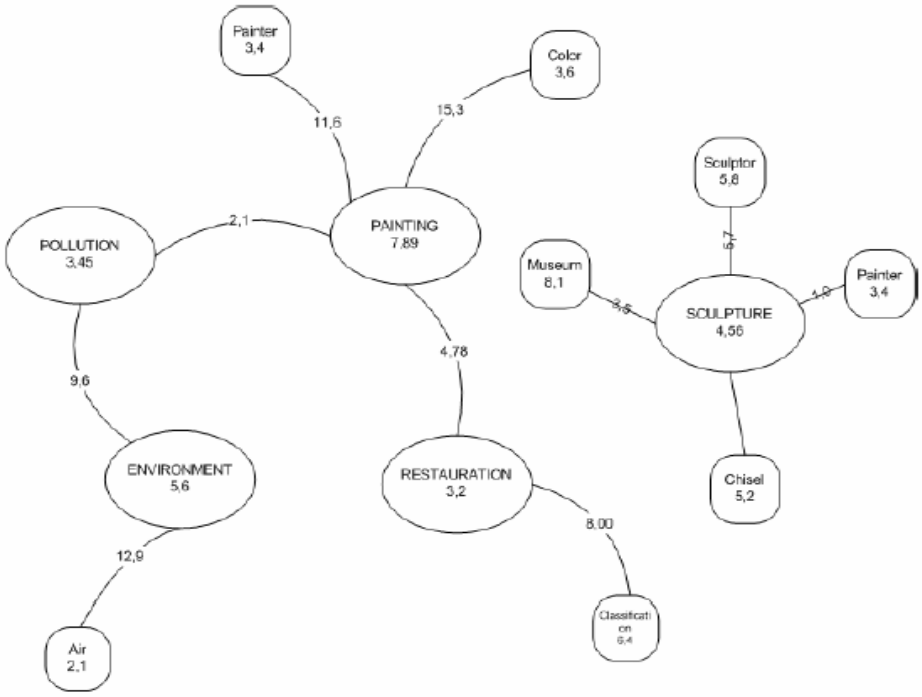


Figure 3.2 — Un extrait d'un profil utilisateur sémantique

Une approche similaire a été utilisée par le système InfoWeb [67]. Initialement, chaque réseau sémantique contient une collection de nœuds unitaires dans lesquels chaque nœud représente un concept. Les nœuds du concept, appelés *planètes*, contiennent un vecteur unique de termes pondérés. Lorsque de nouvelles informations sur l'utilisateur sont collectées, le profil est enrichi en intégrant des termes pondérés dans les concepts correspondants. Ces termes sont stockés dans des nœuds auxiliaires, appelés *satellites* qui sont liés aux nœuds concepts (planètes) associés. Des liens sont également ajoutés entre les planètes représentant des associations entre les concepts. La figure 3.2 montre un exemple extrait d'un modèle d'utilisateur basé sur cette représentation.

Cette représentation a été prolongée dans WIFS [120], une interface de filtrage d'information pour personnaliser les résultats du moteur de recherche d'AltaVista [3]. Dans ce système, le profil de l'utilisateur est représenté par trois composantes : un entête, intégrant les données personnelles de l'utilisateur, un ensemble de stéréotypes, et une liste de centres d'intérêts. Un stéréotype, comporte un ensemble de centres d'intérêts, représenté par une classe d'informations. Chaque classe contient trois champs : *domaine*, *matière*, et *poids*.

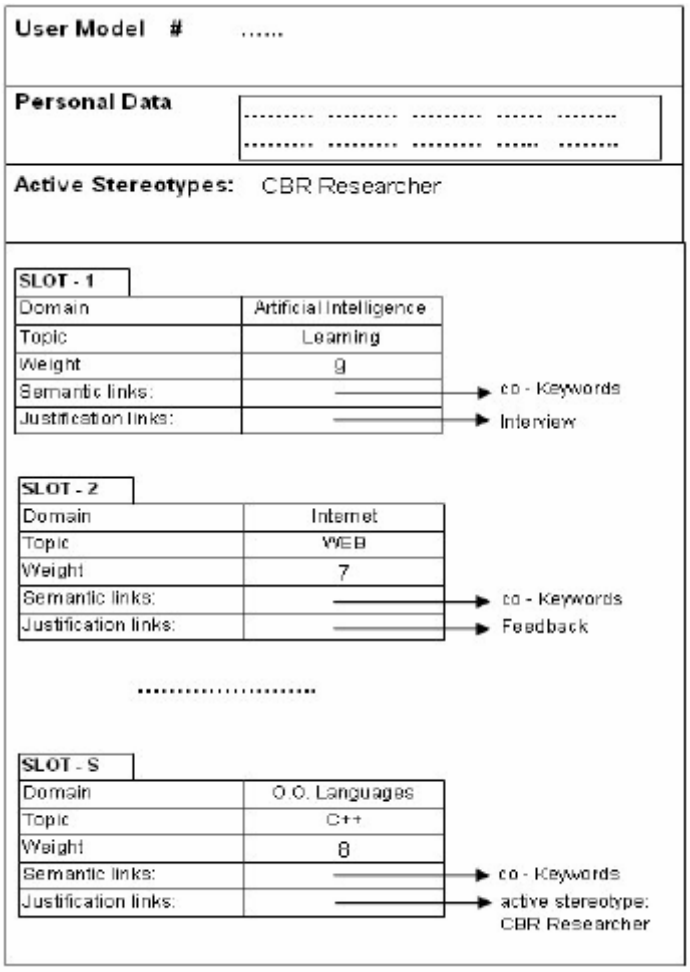


Figure 3.3 — Un extrait du profil sémantique de WIFS

Le domaine identifie un centre d'intérêt de l'utilisateur, la matière contient le terme spéci-

fique employé par l'utilisateur pour décrire son centre d'intérêt, et le poids indique le degré d'intérêt de l'utilisateur pour ce centre. Le profil utilisateur est représenté par un groupe de classes contenant les liens sémantiques et les liens de justification pour l'acquisition des données. La figure 3.3 montre un extrait de profil basé sur cette représentation. Les liens sémantiques incluent les listes de mots-clés co-occurrents dans le document lié à la classe et ayant un degré de similarité avec la matière. De ce fait, le profil est vu comme un ensemble de réseaux sémantiques, pour lesquels une classe est une planète et les liens sémantiques sont les satellites.

3.2.3 Représentation conceptuelle

La représentation du profil met en évidence, dans cette approche, les relations sémantiques entre les informations le contenant. Suivant une direction proposée dans un contexte plus général par Huhn [78], cette représentation offre une alternative intéressante à l'approche précédente.

En effet, les travaux actuels tendent à représenter le profil sous forme d'une ontologie de concepts personnels en se basant sur les connaissances contenues dans les ontologies plutôt que de construire les profils d'utilisateur seulement à partir des documents collectés de son interaction [35]. La représentation est essentiellement basée sur l'utilisation d'ontologies [131; 65; 36; 11; 113; 73] ou des réseaux probabilistes [111; 190]. Dans ce type d'approche les liens entre les concepts sont explicitement induits de l'ontologie et le profil résultant inclura des relations informationnelles plus diverses et spécifiques.

La représentation conceptuelle est semblable à la représentation sémantique, dans le sens, où elle représente les centres d'intérêts de l'utilisateur par un réseau de nœuds conceptuels. Cependant, dans l'approche conceptuelle, les nœuds correspondent à des domaines abstraits représentant les centres d'intérêts de l'utilisateur, contrairement à l'approche sémantique où les centres d'intérêts sont représentés par un mot spécifique ou ensemble de mots relatifs. La représentation conceptuelle peut également être assimilée à une approche ensembliste (vectorielle) du fait que les domaines sont souvent représentés comme des vecteurs de termes pondérés. Néanmoins, les termes de ces vecteurs sont regroupés pour former un domaine spécifique et non de simples mots-clés.

De l'association des centres d'intérêts de l'utilisateur aux concepts des domaines de l'ontologie, on obtient un profil représenté sous forme d'une hiérarchie de concepts. Les documents qui intéressent l'utilisateur sont ensuite classifiés dans ces concepts et l'intérêt de l'utilisateur pour de tels concepts est enregistré. Plusieurs mécanismes peuvent être appliqués pour exprimer le degré d'intérêt de l'utilisateur pour chaque domaine. Néanmoins la technique la plus utilisée est l'affectation d'une valeur numérique ou d'un poids à chaque domaine [65].

On trouve dans la littérature plusieurs types de structures hiérarchiques et ressources sémantiques, les plus simples sont construits sur la base d'une taxonomie de concepts ou d'un thésaurus de référence. A titre d'exemple, les systèmes [70; 96] utilisent l'ontologie *Sensus*, une taxonomie d'approximativement 70.000 noeuds, et un sous-ensemble de l'annuaire de Yahoo ! [107; 196] en tant que hiérarchie de référence.

On trouve également, ODP[1] qui est une hiérarchie de concepts open source au format *RDF* largement adoptée par de nombreux systèmes appliquant l'approche conceptuelle tels que OBIWAN (Ontology Based Informing Web Agent Navigation) [138], *Personae* [178] et *Outride* [136].

L'utilisation d'ODP comme source conceptuelle diffère d'un système à un autre. Ainsi, dans OBIWAN les profils sont représentés en utilisant 1.869 concepts des trois principaux niveaux de la hiérarchie de concepts d'ODP [35]. Avec l'élargissement du contenu d'ODP, le nombre de concepts utilisés est augmenté à environ 2.991 concepts extraits également des trois principaux niveaux [181]. En outre, le système *Personae* [178] ne se limite pas aux premiers niveaux d'ODP, mais exploite différents concepts extraits de différents niveaux de la hiérarchie ODP. Il établit ainsi des profils utilisateurs plus spécifiques. Le système *Personae* gère ainsi des profils de moyenne taille. Quand au système *Outride* [136], dont les profils utilisateurs sont également plus petits que ceux utilisés dans OBIWAN, exploite seulement 1.000 concepts d'ODP.

L'organisation sous forme hiérarchique des concepts du profil est adaptée différemment pour chaque utilisateur afin de mieux exprimer les caractéristiques de généralisation/spécification inhérente à ce type de structure [17]. Les niveaux de la hiérarchie peuvent être statiques et fixes [181], ou changer dynamiquement selon les degrés d'intérêt de l'utilisateur attribués à chaque concept [38].

Dans un contexte aussi dynamique et volumineux que le *web*, la représentation du profil basée sur les ontologies engendre certains problèmes liés à l'hétérogénéité et la diversité des centres d'intérêts de l'utilisateur. En dépit du fait que ces profils peuvent contenir un nombre considérablement élevé de concepts, ces concepts n'englobent que partiellement le nombre potentiellement infini des centres d'intérêts spécifiques de chaque utilisateur. Par exemple, Yahoo ! peut représenter le concept de *base-ball* à l'intérieur de celui du *sports* mais ne pas représenter un intérêt plus spécifique pour une équipe ou un joueur célèbre (ou pas) de base-ball. En outre, les ontologies imposent leur organisation des concepts aux profils d'utilisateur qui ne sont pas nécessairement en correspondance avec les perceptions de l'utilisateur. D'ailleurs, les utilisateurs peuvent avoir différentes perceptions pour un même concept, ce qui engendre des représentations imprécises de l'utilisateur [68].

[1] The Open Directory Project (ODP), *http ://dmoz.org*

3.2.4 Représentation multidimensionnelle

Les utilisateurs sont divers et complexes : ils sont caractérisés par des modèles cognitifs différents et font partie d'une communauté. Ils effectuent des tâches multiples ayant des buts différents. Ils ont également des activités simultanées de recherche, interactives et connexes à d'autres entités dans un domaine donné. On constate ainsi que les informations caractérisant un utilisateur ne sont pas factuelles mais multidisciplinaire. Cependant, cette diversité n'est généralement pas fidèlement représentée par les modèles de profil présentés précédemment.

Inscrit dans une réflexion globale sur la personnalisation de l'information, une autre représentation possible du profil est la représentation multidimensionnelle. Cette représentation a pour objectif de capturer toutes ces caractéristiques informationnelles de l'utilisateur.

Différents travaux ont abordé cet aspect sans le couvrir dans son ensemble. Ainsi, les propositions de standards P3P [189] pour la sécurisation des profils ont défini des classes distinguant les **attributs démographiques** des utilisateurs (*identité, données personnelles*), les **attributs professionnels** (*employeur, adresse, type*) et les **attributs de comportement** (*trace de navigation*).

Une autre proposition faite par Amato [4] consiste à représenter le contenu du profil utilisateur par un modèle structuré de dimensions (ou catégories) prédéfinies : *catégorie de données personnelles*, *catégorie de données de la source*, *catégorie de données de livraison*, *catégorie de données de comportement* et *catégorie de données de sécurité*. L'auteur a proposé ce modèle dans le cadre du développement d'un service avancé de librairie digitale (recherche et livraison personnalisées de l'information sur le *web*) : le système EUROgatherer.

Dans ce même cadre, Kostadinov [102] a poursuivi cette classification en proposant un ensemble de dimensions ouvertes, pouvant contenir la plupart des informations susceptibles de caractériser l'utilisateur. Dans sa représentation il distingue principalement huit dimensions décrites brièvement dans ce qui suit :

* *Les données personnelles*

  Les données personnelles sont la partie statique du profil. Elles comprennent l'identité civile de l'utilisateur (nom, prénom, numéro de sécurité sociale, etc.) ainsi que des données démographiques (âge, genre, adresse, situation familiale, nombre d'enfants, etc.)
* *Le centre d'intérêt*

  Le centre d'intérêt exprime le domaine d'expertise de l'utilisateur. Il peut être défini par un ensemble de mots clés ou un ensemble d'expressions logiques (requêtes).
* *L'ontologie du domaine*

  L'ontologie du domaine complète la définition du centre d'intérêt en explicitant la sémantique de certains termes ou de certains opérateurs employés par l'utilisateur dans son profil ou dans ses requêtes.

* *La qualité attendue*

  La qualité est un des facteurs clés de la personnalisation, elle permet d'exprimer des préférences extrinsèques comme l'origine de l'information, sa précision, sa fraîcheur, sa durée de validité, le temps nécessaire pour la produire ou la crédibilité de sa source. Les attributs de cette dimension expriment la qualité attendue ou espérée par l'utilisateur.

* *La customisation*

  La customisation concerne d'abord tout ce qui est lié aux modalités de présentation des résultats en fonction de la plateforme, de la nature et du volume des informations délivrées, des préférences esthétiques ou visuelles de l'utilisateur.

* *La sécurité*

  La sécurité est une dimension fondamentale du profil. Elle peut concerner les données que l'on interroge ou modifie, les informations que l'on calcule, les requêtes utilisateurs elles-mêmes ou les autres dimensions du profil. La sécurité du processus exprime la volonté de l'utilisateur à cacher un traitement qu'il effectue.

* *Le retour de préférences*

  On désigne par ces termes ce qu'on appelle communément le « feedback » de l'utilisateur. Cette dimension regroupe l'ensemble des informations collectées sur l'utilisateur.

* *Les informations diverses*

  Certaines applications demandent des informations spécifiques ne pouvant être incluses dans aucune des dimensions précédentes comme par exemple la bande passante attribuée au gestionnaire du profil. Pour cette raison l'utilisateur a la possibilité de rajouter ce type de préférences dans la partie divers du profil et de décrire leurs utilisations.

Pour une application donnée, un utilisateur n'a pas besoin de toutes les dimensions ou sous dimensions ni de toutes les informations caractérisant une dimension. Un profil donné est donc une instanciation partielle de ce méta modèle en fonction des besoins de l'utilisateur, du type d'application et de l'environnement d'exécution de cette application.

Par la suite, ce travail a été prolongé dans le cadre du projet Accès Personnalisé à des Masses de Données (APMD)[2]. Ce projet a pour objectif de mener une réflexion globale sur la personnalisation de l'information dans un environnement à grande échelle aussi bien dans un environnement de RI que de BD. Les travaux menés dans ce cadre par Kostadinov [103] ont abouti à la proposition d'un modèle utilisateur générique englobant les notions de profil, de préférence et de contexte. Un profil utilisateur représente l'ensemble des informations décrivant l'utilisateur. Un contexte représente les données décrivant l'environnement d'interaction entre un utilisateur et un système. Une préférence est une expression permettant de hiérarchiser l'importance des informations dans un profil ou un contexte.

[2]ACI Masses de Données : Projet MD-33 Accès Personnalise à des Masses de Données. *http ://apmd.prism.uvsq.fr/*

3.3 Approches de construction du profil utilisateur

La construction du profil traduit un processus qui permet d'instancier sa représentation. L'approche de construction dépend fortement de la représentation choisie pour le profil utilisateur : les techniques utilisées par les systèmes différents selon qu'ils représentent le profil par un (des) vecteur(s) de termes ou par des classes (hiérarchiques ou pas). Cependant la démarche de construction commune à tous les systèmes est la suivante : on commence par collecter des informations sur l'utilisateur à partir de sources d'informations diverses, puis on applique des techniques et des algorithmes pour apprendre à partir de ces informations le profil de l'utilisateur. La construction du profil s'effectue donc en deux étapes : (1) l'acquisition et la collecte des données utilisateur ; (2) puis la construction proprement dite du profil.

3.3.1 Acquisition des données utilisateurs

Cette phase consiste à collecter les informations pertinentes pour instancier le profil de l'utilisateur. Le processus d'acquisition des données de l'utilisateur implique différentes formes de diagnostic ou d'évaluation. Ce processus peut collecter ces informations soit directement à partir la machine de l'utilisateur (côté client) ou à partir de l'application (côté serveur). Ce processus d'acquisition peut être explicite et/ou implicite :

1. l'approche explicite consiste à obtenir les informations directement de l'utilisateur,
2. l'approche implicite, largement motivée par les travaux actuels dans le domaine, implique l'exploitation des données de comportement de l'utilisateur pour inférer son profil.

Nous détaillons ces deux approches dans ce qui suit, puis nous dresserons un bilan comparatif et nous donnerons une synthèse des principaux travaux dans ce domaine.

3.3.1.1 L'acquisition explicite

Cette technique constitue une approche simple pour obtenir des informations sur l'utilisateur. On interroge directement l'utilisateur ou on lui demande par exemple de remplir des formulaires pour collecter les données personnelles et démographiques tels que sa date de naissance, son statut marital, son activité professionnelle et ses centres d'intérêts.

Dans le cadre de l'accès personnalisé à l'information, l'approche explicite est assimilable au feedback explicite, largement utilisé dans les systèmes de filtrage et de reformulation de requête. En effet, l'utilisateur émet directement son jugement d'intérêt en donnant une valeur de pertinence sur une échelle graduée allant du moins intéressant au plus intéressant.

L'acquisition explicite a été largement utilisée dans les systèmes de E-commerce pour personnaliser l'interface des sites *web* en fonction des préférences des utilisateurs.

Ces systèmes peuvent être considérés comme les premières approches de personnalisation. A titre d'exemple, MyYahoo ! [196] demande explicitement à l'utilisateur de fournir les informations personnelles qui seront stockées pour créer le profil utilisateur. Le contenu du site *web* MyYahoo ! est alors automatiquement organisé sur la base des préférences de l'utilisateur contenue dans ce profil, pour ne présenter que les informations intéressants l'utilisateur. D'autres systèmes d'accès personnalisé, plus élaborés, basés sur le feedback se sont concentrés sur l'activité de navigation de l'utilisateur. Parmi ces systèmes, on peut citer Syskill & Webert [134] et WAWA (Wisconsin Adaptive Web Assistant) [158; 157].

Ce procédé d'acquisition direct des données est sûr et souvent utilisé pour déclencher la collecte implicite d'informations. Cependant cette méthode induit un désintéressement et l'abandon de l'utilisateur, ce qui en résulte une détérioration de l'efficacité du système de recherche. Ceci est dû à la surcharge cognitive de l'utilisateur engendré par l'obtention répétitive et fréquente du jugement de pertinence. De plus la construction du profil dépend fortement du degré d'implication de l'utilisateur : si l'utilisateur ne fournit pas volontairement les informations, aucun profil ne sera construit.

3.3.1.2 L'acquisition implicite

L'acquisition implicite ou « *feedback implicite* » consiste à collecter les données de l'utilisateur, en observant son comportement et en scrutant son activité. L'activité peut correspondre à :

* l'utilisation de moteur de recherche : requêtes et documents sélectionnés,
* la navigation sur le *web* : pages *web* consultées, liens sélectionnés,
* diverses applications utilisées dans le contexte de sa recherche : les applications du bureau (tels que les produits MSOffice), les outils de messagerie électronique, les éditeurs de texte, les fichiers logs,
* consultation de bases de données ou des bases documentaires.

Le principal avantage de cette approche est qu'elle ne nécessite aucune implication directe de l'utilisateur, ni de temps passé à émettre des jugements, ni un effort d'attention particulier lors de sa recherche. En effet, toute interaction de l'utilisateur avec le système est considérée comme une estimation de son jugement d'intérêts, tel que l'explique [133] qui stipulent que :

"*Implicit feedback (method) may bear only an indirect relationship to the user's assessment of the usefulness of any individual document*"

Basé sur la durée de l'interaction avec le système, le procédé d'acquisition implicite du profil d'utilisateur peut être classifié comme suit [141] :

* **Un modèle peu profond** basé sur l'observation du comportement d'interaction relativement à court terme avec un système ; il ne tient pas compte des interactions de l'utilisateur avec le système durant les sessions précédentes.
* **Un modèle profond** qui observe le comportement de l'utilisateur durant son interaction à long terme avec un système. Il tient compte de l'historique entier des interac-

tions de l'utilisateur avec le système.

La difficulté de ce type de techniques est la définition d'un processus d'interprétation du comportement observé dans un contexte d'application spécifique. Le processus doit être capable d'inférer les prétentions (intention, centres d'intérêts, préférences) de l'utilisateur. Le plus souvent ces prétentions sont incertaines et doivent être validées par l'utilisateur ou par des observations du comportement à plus long terme.

Nous exposons dans ce qui suit les principaux comportements observables de l'utilisateur lors de son interaction avec les SRIP. Puis, nous abordons les différentes sources d'information pour la construction du profil utilisateur.

1. **Comportements observables de l'utilisateur**

   Un comportement observable de l'utilisateur est l'ensemble des actions qu'il effectue face aux résultats fournis par le système d'accès à l'information. L'interprétation de ce comportement sera effectuée grâce à un groupe d'indicateurs implicites.

   Se basant sur les travaux de [132], [133] proposent une catégorisation des comportements observables de l'utilisateur. Ils les classent en fonction de **la catégorie de comportement** et **des unités élémentaires manipulées**. La catégorie de comportement (*examiner, sélectionner, mettre en référence et annoter*), se rapporte au but fondamental du comportement observé. L'unité élémentaire manipulée (*segment, objet et classe*) se rapporte à la plus petite unité informationnelle manipulée par l'utilisateur à ce moment. Kelly [92] a ajouté une cinquième catégorie de comportement, "*Création*", aux quatre catégorisations d'Oard et de Kim. Cette catégorie décrit les comportements que l'utilisateur lors de la création de nouvelles unités informationnelles comme par exemple l'écriture d'un papier. L'ensemble de ces classifications est regroupé, ainsi que les unités observables associées, dans le tableau 3.1.

   L'interprétation de ces comportements nécessite l'utilisation d'indicateurs traduisant l'intérêt de l'utilisateur pour les données manipulées. En effet, pour obtenir du comportement de l'utilisateur son intérêt pour un résultat de recherche, il est nécessaire d'utiliser des indicateurs traduisant cet intérêt. La définition et le choix de ces indicateurs constituent la principale problématique pour la mise en œuvre de ce type de techniques. La majorité des travaux de recherche sur le feedback implicite se sont focalisés sur la définition d'indicateurs permettant d'obtenir de bonne prédiction sur la pertinence du comportement observé de l'utilisateur. Pour atteindre cet objectif, ils ont effectué à chaque expérimentation des comparaisons entre les résultats obtenus par feedback explicite (supposé plus fiable par l'intervention de l'utilisateur) et ceux obtenus du feedback implicite. Le détail de ces travaux ainsi que la typologie des indicateurs de comportement pourra être consulté en annexe A.

Catégories des comportements \ L'unité sur laquelle se porte l'observation	Segment	objet	classe
Examiner	Regarder Ecouter Défiler Trouver Soumettre une requête	Sélectionner	Naviguer
Retenir	Imprimer	Marquer Sauvegarder Supprimer Envoyer un mail	
Référencer	Copier - Coller	Répondre Ajouter un lien Citer	
Annoter	Marquer	Juger Publier	Organiser
Créer	Taper, Editer	Autre	

Tableau 3.1 — La catégorisation du comportement utilisateur selon [133; 91]

2. **Sources d'information pour l'acquisition implicite**

   Différentes sources d'information sont utilisées par les systèmes d'accès pour construire le profil de l'utilisateur. Les informations sont principalement issues des interactions utilisateur-système lors de l'accès de ce dernier. On distingue principalement deux types de sources d'information :

   (a) *L'historique des interactions* (ou de navigation) est obtenu généralement par le système en scrutant les actions effectuées sur la page *web* tel que les annotations, le référencement et sauvegarde. Il peut contenir les URLs des pages *web* consultées par l'utilisateur, ainsi que la date et le temps de chaque consultation dans le but d'extraire des données statistiques sur le comportement et préférences de l'utilisateur (tel que le nombre de consultation et le temps de lecture).

   La technique la plus communément utilisée par la majorité des systèmes se base sur l'utilisation d'un serveur proxy pour observer les interactions de l'utilisateur avec le système. Le principal avantage est qu'elle ne nécessite qu'une implication minimale de l'utilisateur et aucune installation de logiciel. Cependant, peu de profils peuvent être construits car l'accès s'effectue souvent de manière anonyme par différents utilisateurs sur différents emplacements géographiques.

   Une alternative possible est de s'enregistrer sur le système. De ce fait, le profil sera alimenté par son historique à chaque connexion. Peu importe son emplacement géographique le système lui fournira un accès personnalisé en fonction de son

profil [181; 9; 138].

D'autre part, les approches de personnalisation peuvent employer des agents (directement installés sur la machine de l'utilisateur, intégrés de manière autonome dans le système d'accès ou dans les navigateurs *Web* standards) pour capturer l'ensemble des activités effectuées par l'utilisateur parallèlement à sa recherche.

Letizia [110] est l'un des premiers systèmes de personnalisation ayant exploité les *feedback* implicites à l'aide d'agents. Basé sur les pages précédemment visitées et annotées par l'utilisateur, il suggère des liens intéressant l'utilisateur sur sa page courante. Plusieurs autres systèmes d'aide à la navigation ont suivi tels que WebMate [39], Vistabar [118], et Personal WebWatcher [125]. D'autres approches de personnalisation mettent en place des toolbars pour aider l'utilisateur à organiser l'information contenue sur son ordinateur. Parmi ces systèmes on trouve Seruku Toolbar [156] et SurfSaver [170].

(b) *L'historique de recherche* constitue une source d'information pour la construction du profil de l'utilisateur dans le cadre d'une activité de RI. Il contient des informations issues des requêtes de l'utilisateur (requêtes, documents résultants, dates et durée des requêtes). Le SRIP peut également exploiter les URLs et les résumés ou les documents retrouvés par le système en réponse à la requête, ayant été sélectionnés par l'utilisateur. Le principal avantage de cette ressource est qu'elle ne nécessite aucune installation logicielle. En effet, le SRIP collecte directement l'historique de recherche de l'utilisateur à partir de son interaction avec les résultats de recherche. Si le système nécessite l'enregistrement de l'utilisateur, son profil sera alimenté à chaque connexion, et un même profil sera exploité à chaque nouvelle recherche. Cependant, le risque potentiel est que les informations collectées de l'utilisateur seront issues uniquement de l'historique des résultats de recherche retournées par le SRIP lui même. Ce qui constitue une source d'information moins générale que dans les approches précédentes. Néanmoins, plusieurs projets [114; 165] ont obtenu de très bonnes performances de recherche personnalisée basées sur des profils d'utilisateur construits à partir de l'historique de recherche.

Dans le cas où l'accès n'est pas limité au *web*, le système inclut des sources d'information plus générales telles que les bases de données utilisées par l'utilisateur, ses documents personnels disponibles dans son environnement de travail. Cette approche est mise en application par des outils tels que Google Desktop [55] et SIS (Stuff I've Seen) [57]. En effet, le profil de l'utilisateur est alimenté implicitement à partir de bases de données et de ses fichiers personnels. Dans la même perspective le projet Haystack [1; 40] fournit une infrastructure pour créer un environnement personnalisé : une base de données pour stocker tous les documents de l'utilisateur, un système de gestion de base de données, et un module d'apprentissage du profil utilisateur.

3. **Synthèse des techniques implicites**

Le tableau 3.2 résume les différentes techniques d'acquisition implicites. Il présente en

particulier le type de source d'information collectée, l'avantage et l'inconvénient de la technique et son applicabilité dans le système.

Technique	**Source**	**Applicabilité**	**Avantage-Inconvénient**	**Système**
Historique du navigateur	Historique de recherche	Sites *web*	(+) Aucune installation (-) sauf mise à jour	OBIWAN [138]
Serveur proxy	Hisorique de recherche	Sites *web*	(+) Utilisation de navigateur standard et (-) d'un serveur proxy	OBIWAN [138] [181] [9]
Agent de recherche	Historique de recherche	Accès personnalisé	(+) L'agent collecte tout activité utilisateur (-) Installation de l'agent et utilisation de l'application spécifique lors de la recherche	Letizia [110] WebMate [39] Vistabar [118] Personal WebWatcher [125]
Agent Desktop	Toute activité utilisateur	Accès personnalisé	(+) Tous les fichiers et activité de l'utilisateur (-) Installation de l'agent	Seruku [156] Surfsaver [170] Haystack [1] [40] Google Desktop [55] Stuff I've Seen [57]
Les logs de navigation (Cookies)	Toute activité utilisateur	Le site *web*	(+) Information collaborative de plusieurs utilisateurs (-) Information minimale issue d'un seul site *web*	Mobasher [126]
Les logs des requêtes (Cookies)	Résultats de recherche	Moteur de recherche	(+) Information issue d'un même moteur (-) Information minimale issue d'un seul site*web* (-) Activation des Cookies à chaque connexion	Misearch [160] [114]

Tableau 3.2 — Synthèse des techniques d'acquisition implicite

3.3.1.3 Discussion : acquisition explicite vs. acquisition implicite

L'acquisition implicite des données utilisateur pour la construction de profil utilisateur n'a été investie que récemment par la communauté de la RI.

En 2000, Quiroga et Mostafa [140] ont comparé les performances du feedback explicite et implicite en analysant les résultats d'un système de filtrage d'information avec 18 utilisateurs sur une collection de test de 6.000 Mo disques de santé médicale classifiés en 15 domaines différents. Chaque utilisateur a utilisé le système durant 15 sessions de recherche. Ils ont obtenu une amélioration de la précision d'une valeur approximative à 68% lors de l'utilisation de profils utilisateurs construits à partir de la combinaison des deux approches de feedback. En outre, les résultats basés uniquement sur le feedback explicite ont produit une précision maximale autour de 63% et de 58% comparativement à ceux basés uniquement sur l'implicite. Ces différences s'avèrent statistiquement significatives. Ces expérimentations suggèrent que l'utilisation d'un profil explicitement construit ou un profil établi par la combinaison des deux approches produit de meilleurs résultats qu'une recherche basée sur un profil uniquement construit de manière implicite.

Cependant, contrairement aux précédents résultats, [191] n'ont pas trouvé des différences significatives entre les profils construits implicitement et explicitement. En effet, ils ont comparé deux systèmes de recherche sur le *web*, chacun basé sur le feedback explicite et implicite respectivement. Les expérimentations effectuées ont été menées par 16 utilisateurs ayant pour objectif de retrouver l'information sur le *web* en réponse à des requêtes spécifiques portant sur quatre domaines d'intérêts. L'accomplissement réussi de la recherche, la durée, ainsi que le nombre de pages résultant consultés pour chaque recherche ont été choisis comme métriques d'évaluation des performances des systèmes. Les utilisateurs avec des profils implicitement construits ont consulté approximativement $3,3$ pages retrouvées pour chaque recherche, ce qui est d'avantage plus élevé que les $2,5$ pages consultées dans le cas des profils explicites. Les auteurs ont conclu que les approches explicites et implicites étaient similaires car ces différences n'étaient statistiquement pas significatives.

Plus récemment, [180] ont évalué diverses sources d'informations pour la construction du profil utilisateur tels que les pages *web* visitées, les *messages* échangés et l'ensemble des documents stockés sur la machine de l'utilisateur. Ils ont testés plusieurs profils utilisateurs construits à partir de différentes collections de tests issues par exemple uniquement des documents récemment enregistrés, des pages *web* uniquement, et combinaison des différentes sources. En outre, ils ont construit deux profils utilisateur à partir de l'historique de recherche de l'ensemble des requêtes (préalablement soumises) et à partir de l'ensemble des domaines d'intérêts consultés lors de la navigation de l'utilisateur. Ils ont constaté que les performances de recherche augmentent corrélativement avec la quantité d'information utilisée pour construire le profil utilisateur.

Plus particulièrement, ils ont obtenus l'ordre d'importance des profils suivant :

1. le profil construit à partir de l'indexation de l'ensemble du bureau de l'utilisateur (l'ensemble de toutes les informations créées, copiées, ou vues par l'utilisateur)
2. le profil établi uniquement à partir de l'information récemment consultée,
3. le profil basé uniquement sur des pages *web*,
4. le profil le moins précis a été obtenu uniquement des requêtes soumises par l'utilisateur, néanmoins les résultats de recherche obtenus par ce profil ont surpassé une recherche classique (non personnalisée)

Ces résultats peuvent dans un sens valider l'approche implicite par rapport à l'approche explicite.

Ces différentes études, prises dans leur ensemble, suggèrent que les méthodes implicites sont sensiblement plus performantes que l'approche explicite pour la construction des profils utilisateur. Le principal avantage de l'approche implicite est qu'elle est totalement transparente dans le processus d'acquisition des données pour l'utilisateur. De plus, le profil peut être mis à jour plus fréquemment que lors d'une construction explicite. Ces mises à jour permettent la coordination de l'évolution du profil de l'utilisateur à long terme. Néanmoins, l'inconvénient inhérent aux techniques implicites est qu'elles sont incertaines du fait de l'incertitude sous jacente aux comportements des utilisateurs et que les prédictions issues de l'observation de ces comportements sont difficilement quantifiables.

3.3.2 Techniques de construction

Le processus de construction consiste à organiser et extraire les éléments qui constituent le profil à partir des données de l'utilisateur collectées lors de l'étape précédente, selon le modèle de représentation du profil utilisateur. La construction s'appuie sur différentes techniques selon la représentation de profil utilisateur. On distingue trois principales techniques, détaillées dans les paragraphes suivants : l'*extraction des termes*, l'*extraction de réseaux de termes* et l'*extraction de concepts*.

3.3.2.1 Extraction d'ensemble de termes

L'extraction des termes est une technique largement utilisée par la majorité des systèmes de personnalisation. Elle est basée sur des techniques d'analyse statistique de mots clés.

L'idée principale consiste à analyser le contenu des documents utilisateur et d'en extraire des mots clés significatifs qui décrivent son contenu. Ces termes constituent les données d'entrée pour l'algorithme d'apprentissage du profil. Dans le cas où le profil contient simplement que des mots-clé, ces termes vont être regroupés en paquets selon leur degré de similarité pour former les centres d'intérêts. Dans le cadre d'une approche vectorielle, les termes vont être pondérés pour former des vecteurs de termes représentant les centres d'intérêts. Le poids attribué à chaque mot clé permet de traduire son degré d'importance dans le profil. La fonction

de pondération appliquée par la majorité des systèmes est issue du schéma $TF*IDF$ [151]. Le nombre de termes extraits est souvent fixé selon un seuil de pondération de sorte que seuls les termes dépassant cette valeur contribuent à la construction du profil. Ceci permet d'obtenir des profils plus concis et plus représentatifs des centres d'intérêts de l'utilisateur.

Parmi les systèmes appliquant cette approche de construction, on peut citer les systèmes *WebMate* [39] et *Alipes* [194]. Dans le cadre du système *WebMate* [39] le profil est représenté par des vecteurs multiples, un par domaine d'intérêts. Le principe de base est d'associer les vecteurs de documents, collectés explicitement des pages *web* consultées par l'utilisateur, à des centres d'intérêts similaires. La formule de pondération de chaque terme t_i d'un vecteur document d est donnée par la formule du $TF*IDF$ suivante :

$$w_{t_i,d} = TF(t_i,d)*IDF(t_i) \tag{3.1}$$

Tel que $IDF(t_i) = \log \frac{|D|}{df(t_i)}$, où $|D|$ le nombre total de documents et $df(t_i)$ est la fréquence d'apparition du terme dans le document, $TF(t_i,d)$ est la fréquence du terme t_i dans le document d.

Le système se base sur un nombre fixe de N centres d'intérêts par profil. Heuristiquement, le nombre N a été fixé à 10. Le profil est représenté par l'ensemble V des vecteurs de centre d'intérêts (initialement ($|V| = 0$)). Chaque vecteur contient M termes. Les vecteurs des N premiers documents, jugés pertinents, sont utilisés pour représenter les N centres d'intérêts. L'apprentissage se déroule de la manière suivante :

1. Prétraitement des données jugées explicitement « intéressantes ». Chaque document est indexé par identification des balises *title*($< TITLE >$), *head1*($< H1 >$), *head2*($< H2 >$), *head3*($< H3 >$) ;
2. Extraire le vecteur de termes de chaque document $d^{(i)}$, noté V_i, en appliquant la formule 3.1 ;
3. Si $|V| < N$ ($|V|$ est le nombre de vecteurs dans l'ensemble V du profil) alors $V \Leftarrow V \cup V_i$
4. Sinon, calculer le cosinus de l'angle entre les vecteurs : $sim(V_j,V_k) = \frac{V_j \bullet V_k}{|V_j| \times |V_k|}$
5. Combiner les vecteurs de plus grande similarité : $V_l = V_l + V_m(l,m) = argmax_{(i,j)}(Sim(V_i,V_j))x,y \in \{1,2,\ldots,n,i\}$
6. Trier les termes du nouveau vecteur par ordre croissant de leurs poids et garder les M premiers termes.

Dans le cas du système *Alipes* [194] le profil est représenté par des vecteurs multiples de termes pour chaque centre d'intérêt, comme nous l'avons déjà mentionné. En effet, chaque centre d'intérêt est représenté par trois vecteurs de mots-clés : à long terme, à court terme (positif) et à court terme (négatif). La construction du profil s'appuie sur les jugements de pertinence (positifs) et de non pertinence (négatifs) de l'utilisateur. Similairement au système *WebMate*, dans *Alipes* le processus d'apprentissage est automatique, à la différence que

la création de nouveaux centres d'intérêts se base sur un seuil de similitude, au lieu d'un nombre prédéterminé de centres d'intérêts. Lorsqu'un vecteur de document est ajouté au profil utilisateur, il est comparé à chacun des trois vecteurs pour chaque centre d'intérêt en utilisant la similitude de cosinus. Si la similitude excède un seuil, le vecteur de document est ajouté au centre d'intérêt le plus proche. Cependant, si aucune similitude n'existe, un nouveau centre d'intérêt est alors créé.

3.3.2.2 Extraction de réseaux de termes

Similairement aux techniques de construction de profils ensemblistes, les termes sont extraits des documents jugés par l'utilisateur. Néanmoins, à la différence des approches précédentes, où les termes forment des vecteurs, les techniques de construction sémantique intègrent ces termes dans un réseau de nœuds.

La construction des profils nécessite l'exploitation de relations préexistantes entre les termes et les concepts, tels que WordNet dans le cas du système SiteIF [168], ou manuellement construites tel que celui effectué par WIFS [120].

Dans les approches élémentaires, chaque utilisateur est représenté par un seul réseau sémantique dans lequel chaque noeud contient un mot-clé unique. Lorsqu'un terme est présent dans le réseau, le poids de son nœud est augmenté ou diminué selon le feedback de l'utilisateur. Si le terme n'apparaît pas dans le réseau, un nouveau nœud est crée. Les poids dans le réseau sont périodiquement réévalués à chaque mise à jour dans le but de modéliser les changements des centres d'intérêts de l'utilisateur à long terme. En outre, les concepts qui ne sont plus d'actualité peuvent être supprimés du réseau.

Dans le cas où les profils construits sont représentés par une collection de centres d'intérêts, où chaque centre est représenté par un réseau sémantique, nous présentons l'approche de construction appliquée par le système WIFS [120]. Ce système effectue un travail préliminaire mené par des experts du domaine. Leur tâche consiste à identifier des ensembles de termes, considérés comme les plus appropriés pour chaque domaine d'intérêt spécifique. En outre, ces experts donnent un ensemble de connaissance de base pour chaque stéréotype, chacune représentant un prototype des besoins en information de l'utilisateur.

Lors de sa première session de recherche, l'utilisateur est interrogé par le système afin d'obtenir un ensemble initial de ses besoins en information. Ces informations sont utilisées pour déterminer le stéréotype (appelé *stéréotype actif*) qui se rapproche le plus de l'utilisateur. Ainsi, le profil de l'utilisateur sera initialisé par les informations issues du questionnaire et les données héritées des stéréotypes actifs.

La construction du profil s'effectue en cinq différents processus, quatre automatiques et un manuel :

1. Premièrement, la mise à jour du centre d'intérêt courant de l'utilisateur, choisi parmi les centres existants, s'effectue par comparaison des termes contenus dans chaque centre d'intérêt du réseau aux termes extraits à partir de la page *web* courante consultée par l'utilisateur et les termes de la requête utilisateur.
2. Deuxièmement, les occurrences des termes de la page *web* existant préalablement dans le réseau pour le centre d'intérêt courant sont utilisées pour modifier le poids (représentant une valeur d'affinité) de l'arc entre le nœud satellite, pour le terme et le nœud planète pour le centre d'intérêt. L'accroissement est proportionnel au degré de pertinence de l'utilisateur, si elle ne dépasse pas un seuil prédéfini, le nœud terme est alors supprimé.
3. Troisièmement, de nouveaux termes extraits à partir de la page *web* sont utilisés pour ajouter de nouveaux nœuds satellites au réseau sémantique correspondant au centre d'intérêt courant. La valeur de pertinence de l'utilisateur pour la page est ensuite utilisée pour pondérer les arcs reliant les nouveaux nœuds satellites au nœud planète représentant le domaine du document.
4. Quatrièmement, si le degré de pertinence de la page *web* excède un seuil, les termes extraits de cette page *web* peuvent être utilisés pour ajouter un nouveau concept dans le réseau.
5. Finalement, les utilisateurs peuvent gérer explicitement leur profil par une modification directe du contenu.

L'uniformité du modèle est maintenue par un système basé sur la justification simple de la maintenance de l'intégrité. Ce système aide à fournir des explications de mise à jour du réseau.

3.3.2.3 Extraction de concepts

Nous allons dans ce qui suit décrire les approches de personnalisation construisant des profils utilisateurs représentés par une hiérarchie de concepts pondérés. Bien que chaque système se base sur une méthodologie de construction différente, ils utilisent tous une taxonomie de concepts de référence comme profil de base (appelé aussi profil général). La plupart des systèmes d'accès utilisent les ontologies comme hiérarchie de référence. Les principales ressources sémantiques utilisées sont ODP[3] un annuaire de concepts hiérarchique open source et la hiérarchie de concepts Yahoo [196]. L'approche de construction s'effectue de manière générale comme suit :

1. identifier les concepts et niveaux de l'ontologie à exploiter. L'objectif étant d'extrait un sous ensemble de concepts représentant le profil général. Dans la plus part des tra-

[3] The Open Directory Project (ODP), *http ://dmoz.org*

vaux la ressource sémantique n'est pas exploité dans sa totalité. Certes, l'utilisation de tous les concepts de la hiérarchie (tel que c'est le cas dans le système *Persona* [178]) permet d'obtenir des profils utilisateurs assez précis, pouvant couvrir un grand nombre de centres d'intérêts. Cependant, la difficulté de cette approche, se situe à juste titre au niveau de la profondeur de la hiérarchie d'ODP et la richesse des concepts. En effet, le profil de l'utilisateur peut devenir très important, contenant trop de concepts proches. Lors de la sélection des documents, le système doit établir des relations de similarité utilisant les concepts du profil de différents niveaux, puis statuer sur le niveau à exploiter pour évaluer les documents candidats. De ce fait, la plus part des travaux extraient qu'un nombre réduit de concepts à partir des premiers niveaux de la hiérarchie [35; 65; 36; 160; 113; 180]. Dans le cas du système PVA, [38] exploitent 55 concepts des trois premiers nivaux de la hiérarchie utilisées par les sites de recherche de [197].

2. extraire les centres d'intérêts de l'utilisateur par analogie aux concepts de l'ontologie. Cette phase correspond à la phase de construction approprement dite. En ce sens où le profil de chaque utilisateur est instacié à partir du profil général (la ressource sémantique) sur la base des informations collectées de l'utilisateur. De manière générale, l'approche consiste en premier à associer à chaque catégorie sémantique de l'ontologie un ensemble de documents représentatifs du concept, puis à projeter sur le profil général, les documents (ou descripteur de documents) issue des différents *feedback* utilisateur pour extraire les concepts représentant le profil de l'utilisateur.

Parmi les différentes approches pour la construction de ces profils conceptuels, nous citons l'exemple de trois systèmes suivants :

Une approche de coloration d'arbre est utilisée dans le système *Persona* [178]. Lorsque l'utilisateur effectue une recherche dans la collection des documents pré-classifiés d'ODP, il doit donner explicitement un jugement de pertinence pour les documents résultants. Le système utilise ce feedback pour mettre à jour son profil. En effet, étant donné que les documents ont été manuellement pré-classifiés dans les concepts d'ODP, le profil est simplement obtenu en prenant en compte le nombre de fois où un concept a été visité, le nombre de jugements de pertinence et de non pertinence associés à chaque nœud, ainsi que l'ensemble des URLs des pages reliées à ce nœud.

Le système ARCH [160] met en œuvre une approche hybride combinant des profils utilisateur basés sur des vecteurs de termes avec une hiérarchie de concepts. Le système collecte à partir du feedback implicite de l'utilisateur les documents ayant suscités l'intérêt de l'utilisateur. Ensuite, ces documents sont classifiés pour identifier leurs domaines d'intérêts et le centroïde de chaque classe est calculé, produisant un vecteur de termes pondérés servant de représentant pour le domaine d'intérêt. Lorsque l'utilisateur soumet une requête, le système identifie le vecteur de centres d'intérêts le plus similaire, puis calcule le degré de similarité entre ce vecteur d'intérêt et les vecteurs de concepts pour identifier le concept approprié.

Des termes de ce concept sont alors utilisés pour reformuler la requête initiale de l'utilisateur.

Dans le cadre du projet OBIWAN, le profil utilisateur est construit en associant les documents collectés de l'observation du comportement de l'utilisateur lors de sa navigation [65; 36; 180] avec les nœuds de l'ontologie. La principale différence entre cette approche et celle du système *Persona* est que le système ne se limite pas à établir les profils utilisateurs des documents pré-classifiés. Toute source textuelle peut être automatiquement classifiée par le système pour trouver les meilleurs concepts appropriés d'ODP et un poids reflétant le degré d'intérêt de l'utilisateur est attribué à ces concepts. Le système construit pour chaque document un descripteur vectoriel, puis il détermine le concept associé de l'ontologie ODP. Par la suite, il applique une fonction de similitude du document avec l'ontologie et classifie le document dans le concept le plus correspondant.

3.3.3 Synthèse des approches de construction

Ce paragraphe présente une synthèse de l'ensemble des approches de construction abordées précédemment. Afin de distinguer entre ces différentes approches, on propose de les regrouper selon l'approche de représentation du profil suivie par les systèmes d'accès personnalisé à l'information. C'est une manière simple et claire de présenter ces approches. Ainsi, on présente dans le tableau 3.3 les approches de construction pour les profils ensemblistes, dans le tableau (3.4) ceux des profils sémantiques et dans le tableau (3.5) ceux des profils conceptuels. Chaque tableau décrit ces approches à travers les sources d'information et la technique utilisée et donne un exemple de système d'accès correspondant.

Représentation	**Sources d'information**	**Technique**	**Exemple**
Vecteur de termes uniques	Pages *web* Feedback implicite positif	Extraction de termes pondérés	Amalthaea [130]
Un vecteur de termes par centres d'intérêts	Pages *web* Liste de centres d'intérêts Feedback implicite (+)	Création de vecteur document Comparer les vecteurs aux centres d'intérêts Fusionner les centres similaires	WebMate [39]
Vecteur de termes multiples par centre d'intérêts	Pages *web* Feedback explicite (+) et (-)	Création de vecteur document Comparer les vecteurs aux centres d'intérêts Ajouter au centre le plus proche	Alipes [194]

Tableau 3.3 — Techniques de construction de profils ensemblistes

Représentation	**Sources d'information**	**Technique**	**Exemple**
Un seul réseau : un nœud par terme	Exemple de documents Pages *web* explicite (+) et (-)	Extraction des termes de plus fort poids Création d'un nœud par terme Relier les nœuds lors d'une Co-occurence du terme dans le document	ifWeb [6]
Un seul réseau : un nœud par concept	Pages *web* Feedback implicite	Extraction des termes de plus fort poids Associer les termes aux concepts WordNet	SiteIF [168]
Un seul réseau : un nœud par concept	Pages *web* Feedback implicite	Extraction de mots Apprentissage des concepts par réseau de neurones	PIN [177]
Un seul réseau : - une planète par concept - un satellite par terme	Ensemble de documents stéréotypes Feedback explicite (+) et (-)	Ajustement directe de l'utilisateur Création de concepts du feedback explicite Ajustement des nœuds et des arcs	InfoWeb [67]
Un réseau par centre d'intérêt : - une planète par intérêt - un satellite par terme	Ensemble de documents stéréotypes Questionnaire utilisateur Feedback explicite	Gestion directe par l'utilisateur Création de nœud concept et nœud terme par des experts mise à jour par raffinement	WIFS [120]

Tableau 3.4 — Techniques de construction de profils connexionnistes

Taxonomie de référence	Sources d'information	Technique	Exemple
Tous les concepts ODP	Feedback explicite (+) des pages *web* pré-classifiées	Coloration d'arbre	*Persona* [178]
Yahoo !	Feedback implicite (+) des pages *web* et le résultat de recherche	Classification	ARCH [160]
97 concepts CORA	Feedback explicite et implicite des articles de recherche pré-classifiés	Coloration d'arbre Propagation de concepts parents	Foxtrot [121]
$\tilde{2}$.000 concepts ODP	Feedback implicite (+) des pages *web* (1) ou historique de recherche (2)	Classification supervisée pour identifier les concepts	(1) OBIWAN [138] (2) Misearch [160]
619 concepts ODP	Feedback implicite (+) de requêtes et résultats de recherche Feedback explicite (+) de catégories	Classification suppervisée pour identifier les concepts Affiner l'apprentissage du classifieur avec des feedbacks	[113]

Tableau 3.5 — Techniques de construction de profils conceptuels

3.4 Approches d'évolution du profil utilisateur

La gestion de l'évolution du profil en fonction des changements des centres d'intérêts de l'utilisateur dans le temps est une caractéristique distinctive des performances des systèmes de personnalisation. La prise en compte de cette évolution peut s'effectuer lors de la modélisation du profil par deux modèles : un modèle à court terme et un modèle à long terme des centres d'intérêts de l'utilisateur [15] et également en appliquant des algorithmes d'évolution génétique [130] et des principes d'évolution liés à la théorie de la vie artificielle [38].

Dans la majorité des systèmes d'accès personnalisé l'évolution du profil a été presque exclusivement limitée à l'addition de nouvelles informations. En d'autres termes, la représentation d'un centre d'intérêt de l'utilisateur est augmentée avec la connaissance de nouvelles données d'interaction, tandis que le plus souvent l'ensemble des centres d'intérêts reste intact. La technique utilisée dans ce cas est inhérente à celle de la construction.

Néanmoins, adapter le profil utilisateur implique des changements au niveau des centres d'intérêts eux mêmes (leurs contenus) qui conduisent éventuellement à la suppression de quelques domaines ou à l'émergence d'autres domaines avec l'augmentation de l'intérêt.

De ce fait, afin de faire face aux changements d'intérêts dynamiques, le système doit non seulement dépister des intérêts dérivant de l'utilisateur afin d'identifier de nouveaux intérêts émergeants, mais également incorporer la connaissance au sujet des expériences d'interactions lors de la recherche. Ainsi, l'évolution consiste à adapter la structure et/ou le contenu du profil aux changements des centres d'intérêts et aux variations des besoins en information de l'utilisateur. A notre connaissance peu de travaux ont abordé cet aspect du profil. En effet, l'évolution n'est prise en considération que dans la mesure où le profil a une existence pérenne, ce qui n'est pas le cas dans la majorité des systèmes car, le plus souvent à chaque connexion de l'utilisateur, un nouveau profil est instancié.

Dans le cas d'une représentation ensembliste, le profil utilisateur évolue en ajoutant de nouveaux vecteurs de termes extraits des documents correspondant aux centres d'intérêts détectés de l'utilisateur. Comme il n'y a souvent pas de dépendance entre les vecteurs, l'ajout d'un nouveau vecteur ne fait qu'augmenter le nombre des centres d'intérêts et non le degré d'importance du domaine.

Dans le cas d'une représentation en classes (hiérarchiques ou non), le processus de mise à jour consiste à mettre à jour graduellement le classifieur. Le système recalcule ainsi pour chaque nouveau document, sa similitude avec les classes déjà existantes :

* Si le document appartient à une classe du profil, le système assigne ce document à cette classe. Il met ainsi à jour le contenu des centres d'intérêts.
* Sinon, le système crée une nouvelle classe traduisant un nouveau centre d'intérêt. Il met ainsi à jour le contenu du profil.

Le système fait évoluer le contenu du profil en associant les nouveaux documents collectés aux classes similaires appropriées. L'adaptation de la structure du profil aux nouvelles classes s'effectue en mettant à jour les relations entre ces classes.

L'utilisation de la théorie de la vie artificielle pour construire et mettre à jour le contenu du profil, est une approche assez novatrice [38]. Elle traduit la notion de cycle de vie d'un centre d'intérêt. En effet, on associe aux classes du profil d'utilisateur une valeur E_i d'énergie, traduisant le degré d'importance d'un centre d'intérêt par rapport à un autre. L'énergie E_i augmente quand les utilisateurs montrent leur intérêt pour les documents de la catégorie i, et elle diminue pour une valeur constante pendant une période. Les catégories (classe d'intérêt) qui ont une grande valeur d'énergie produiraient des sous catégories pour décrire les intérêts de l'utilisateur dans un certain niveau de détail. Respectivement, les catégories qui suscitent peu d'intérêt seront soustraites graduellement et auront finalement tendance à disparaître. Basée sur les valeurs d'énergie des catégories, la structure de la vue personnelle peut être modulée pendant que les intérêts des utilisateurs changent.

3.5 Conclusion

L'approche commune à tous les systèmes d'accès personnalisé à l'information, consiste en premier à modéliser le profil de l'utilisateur, puis à l'intégrer dans le processus d'accès à l'information. Cette démarche n'est cependant pas aussi simple, elle implique une représentation de l'utilisateur par une structure qui permettrait son exploitation par le SRI. Ce chapitre consacré au profil utilisateur, présente une large variété de techniques de modélisation. A partir de nos différentes lectures, nous avons fait ressortir des phases communes dans les approches de modélisation et nous avons identifié : la phase de représentation, la phase de construction (acquisition des données utilisateur, puis apprentissage du profil) ainsi que la phase d'évolution.

A travers cet état de l'art, on remarque que ces approches sont basées sur des modèles dépendant de l'objectif final ou du domaine d'application du système. Plus particulièrement, on note une évolution des approches de représentation des profils utilisateurs, par de simples vecteurs de mot-clé à des représentations plus élaborées et conceptuelles. Ces technologies ont permis d'apporter des solutions pour mieux décrire l'utilisateur selon des aspects cognitifs et comportementaux.

Ce chapitre permet de conclure que la modélisation du profil utilisateur constitue un élément essentiel dans le développement d'un système d'accès personnalisé performant dont l'efficacité dépend fortement de la précision des profils. L'exploitation de ce profil par les systèmes d'accès personnalisé à l'information dépend de plusieurs modèles d'accès que nous détaillons dans le chapitre suivant.

4 L'accès personnalisé à l'information : Modèles d'accès

4.1 Introduction

Dans le chapitre précédent, nous avons discuté les principaux éléments requis pour la personnalisation de l'accès à l'information : techniques et sources d'information pour la modélisation du profil de l'utilisateur. Ce chapitre quant à lui, traite des principaux modèles d'accès personnalisé à l'information et de la validation de ces approches, à travers les campagnes d'évaluation expérimentales.

La section 4.2 dresse un panorama des principaux modèles d'accès personnalisé à l'information et passe en revue quelques systèmes développés durant ces dernières années. Vu le nombre important des systèmes existants, cet état de l'art est nécessairement incomplet, néanmoins il décrit les approches les plus importantes reconnues comme étant de référence dans le domaine de la RI. La section 4.3 présente également une synthèse des approches d'évaluation utilisées. Nous abordons en premier la problématique liée à la mise en place d'une campagne d'évaluation standard et formelle pour l'accès personnalisé. Nous présentons, ensuite, les éléments communs aux approches d'évaluation utilisées dans des travaux de référence dans ce domaine. Ces éléments sont organisés de manière à mettre en évidence les principales étapes d'un protocole d'évaluation de systèmes d'accès personnalisé à l'information.

4.2 Panorama des modèles d'accès personnalisé à l'information

Les systèmes d'accès personnalisé à l'information peuvent être catégorisés selon différents critères : technologie de base, objectif, architecture, etc. Pour notre part, notre objectif est d'extraire à partir d'un ensemble d'approches communes, des méthodologies d'accès pouvant être définies par des modèles formels, en identifiant plusieurs critères de distinction entre les différentes approches de personnalisation. Pour cela, nous menons notre discussion selon les caractéristiques décrites ci-dessous. Cette catégorisation est le résultat d'une ré-

flexion sur l'importance et le rôle central du profil utilisateur dans les différentes phases du processus d'accès.

Ainsi, le premier critère de distinction concerne le cadre applicatif correspondant au service d'accès proposé par le système. Ce critère se définit selon l'objectif du processus d'accès qui peut se distinguer selon deux modes d'accès :

* **Accès sans requête préalable**. Dans ce cas, l'objectif du système est de *recommander* automatiquement des informations personnalisées à l'utilisateur, produites par un ensemble de sources dynamiques. La sélection des documents potentiellement intéressants s'effectue en tenant compte uniquement du profil appris de l'utilisateur. Ce mode d'accès est caractéristique des systèmes de filtrage d'information.
* **Accès avec requête**. L'objectif du système est d'exploiter un ensemble de sources d'évidence extrait du profil de l'utilisateur dans une (ou plusieurs) phase(s) du cycle de vie de la requête. C'est le cadre typique de la RI.

Dans le cas, où l'accès à l'information s'effectue selon le principe des SRI (soumission d'une requête), la personnalisation porte sur l'intégration du profil lors des éventuelles opérations de sélection de sources d'information, de reformulation de requête et de calcul du score de pertinence. Ainsi, le service de personnalisation peut se traduire par l'implémentation de l'une de ces opérations ou d'éventuelles combinaisons possibles. Nous rappelons dans la figure 4.1 les principales phases du processus d'accès personnalisé à l'information selon le cycle de vie de la requête où l'on prend en compte le profil utilisateur, à savoir :

(1) la phase d'exécution de la requête pour la sélection des informations pertinentes,

(2) la phase de présentation des résultats,

(3) la phase d'affinement de la requête.

Ainsi, dans le cadre d'un accès avec requête, on peut affiner le critère de distinction entre les approches d'accès personnalisé à l'information, en se basant sur ces phases. Le facteur de motivation est que plusieurs systèmes peuvent exploiter une même source d'évidence pour apprendre le profil (tel que son contexte applicatif (fichier log de requête, contenu desktop, etc.), son historique d'interaction (feedback implicite, sessions de recherche, etc.) ou des documents pertinents associés (structure des liens, etc.)) cependant ils différent dans les phases d'intégration de ce profil pour fournir un accès personnalisé.

A l'aide de ces différents critères, nous aboutissons à une catégorisation des approches de personnalisation selon les modèles suivants :

* Modèle de recommandation (section 4.2.1),
* Modèle d'appariement personnalisé de l'information (section 4.2.2) ,
* Modèle de ré-ordonnancement des résultats de recherche (section 4.2.3),
* Modèle de reformulation de requête (section 4.2.4),

Il est important de souligner que les systèmes décrits dans chacun de ces modèles se basent essentiellement sur les fondements d'approches théoriques issues de la RI, soutenus par des heuristiques et des algorithmes appropriés (mesures de similarité, classification, réseaux

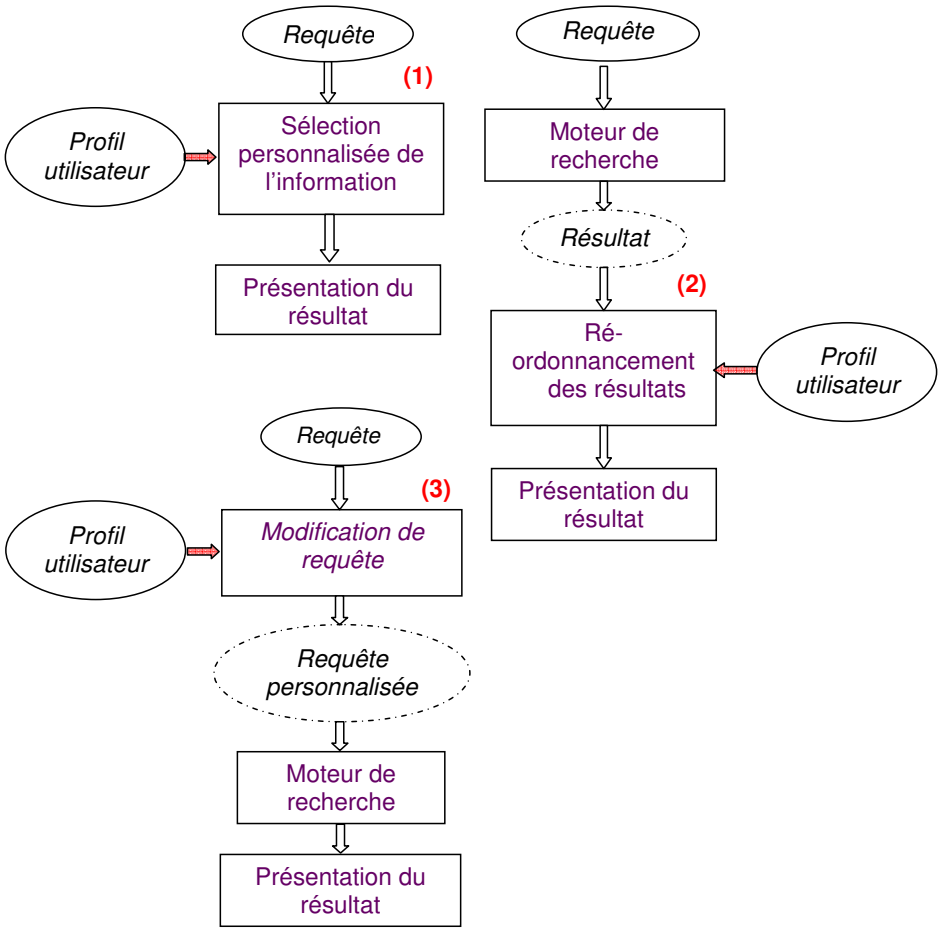


Figure 4.1 — Phases d'intégration du profil utilisateur dans le SRI

Bayésiens, etc.). Il en existe un panel varié. Lors de la présentation de chaque modèle, nous donnons en premier le descriptif global et formel du modèle. Puis, pour chacun des modèles, nous en exposons le détail. Une attention particulière sera portée aux éventuelles particularités caractérisant ces systèmes.

Le premier modèle abordé dans la section 4.2.1 présente les systèmes basés sur la *recommandation* des informations selon le profil de l'utilisateur. La section 4.2.2 présente les modèles dont l'approche de personnalisation s'effectue en tenant compte du profil de l'utilisateur dans le processus de sélection de l'information en personnalisant la fonction de calcul du score de pertinence du document. Le modèle basé sur le ré-ordonnancement des résultats de recherche sera présenté dans la section 4.2.3. Le principe de base est d'affiner la recherche en ne présentant que les résultats en corrélation avec le contenu du profil. Dans la section 4.2.4, nous présentons les approches de personnalisation basées sur *la reformulation de la requête* de l'utilisateur. Ces systèmes réexploitent des approches issues de la RI adaptative pour offrir un accès personnalisé en augmentant la requête par le contenu du profil, afin de mieux cibler la recherche.

4.2.1 Approches de recommandation

La recommandation fut parmi les premières approches suivies pour fournir un accès personnalisé à l'utilisateur. Un modèle de recommandation a pour but de suggérer à l'utilisateur un contenu informationnel susceptible de répondre à ses besoins à long terme en se basant sur un processus de filtrage d'information (FI). Les besoins en informations à long terme de l'utilisateur, que le système essaie de découvrir peuvent être vus comme une requête (ou un ensemble de requêtes).

Le principe de base du filtrage est de traiter les documents provenant de sources dynamiques (News, Email, etc.) et de décider à la volée, si le document correspond ou pas aux besoins en information des utilisateurs, besoins exprimés au travers du concept de profils utilisateurs [14]. Ce processus de recommandation peut s'effectuer selon deux approches de filtrage [117] :

* *Un filtrage par le contenu* (où filtrage cognitif). Il tient compte seulement des contenus du document et du profil.
* *Un filtrage collaboratif* (où filtrage social). Il se base sur les annotations et commentaires attribués par des groupes d'utilisateurs aux documents.

Ces deux approches sont décrites dans ce qui suit.

4.2.1.1 Recommandation basée sur le contenu

Une des premières approches à mettre en œuvre la notion de filtrage d'information a été la DSI (Dissémination Sélective de l'Information) vers les années 60. Elle consiste à filtrer l'information produite par des scientifiques, dans le but de la maintenir continuellement informés des nouveautés relatives à leurs domaines de spécialisation [116]. A partir de 1982, [53] a introduit le principe de filtrage de message email en utilisant des techniques basées sur l'organisation des mail-boxes et nécessitant la coopération des différents usagers. Par la suite la notion de filtrage a été étendue aux articles de presse et articles diffusés sur Internet [62].

Par la suite, cette approche a été principalement développée dans le cadre du *web* : accès par navigation, assimilable à un assistant personnel. L'assistance à la navigation est réalisée soit en aidant l'utilisateur à sélectionner les liens pertinents d'une page, soit en lui soumettant des documents en relation avec le document en cours de lecture, via une exploration automatique des liens.

Letizia [110], l'un des premiers systèmes de personnalisation, s'intéresse essentiellement au service rendu à l'utilisateur, et aux interactions avec ce dernier. Le système accompagne un utilisateur qui explore le *web* en analysant en arrière plan les liens partant des pages consultées. Il utilise pour cela un ensemble d'heuristiques (sauvegarde d'une référence par l'utilisateur, sélection d'un lien, abandon d'une lecture, etc.) pour identifier les documents ayant intéressé l'utilisateur et en déduire ses objectifs. Il explore ensuite dynamiquement

les liens accessibles depuis le document en cours de lecture par l'utilisateur pour évaluer les documents correspondants. Les liens sont enfin ordonnés en estimant leur intérêt pour l'utilisateur. Cette liste est présentée à l'utilisateur lorsque ce dernier demande une recommandation.

En outre, l'accent peut être mis sur les processus utilisés pour apprendre les caractéristiques des documents que l'utilisateur juge intéressants. En ce sens, ***Syskill & Webert*** [134] proposent, en plus de l'analyse des liens présents dans les documents de façon similaire à Letizia, d'utiliser (à la différence de ce dernier) une modélisation de l'utilisateur (un profil) évoluant au cours des recherches à l'aide d'un mécanisme d'apprentissage, dont la description est fournie par l'utilisateur. Initialement, à partir d'un ensemble de documents déjà classés par l'utilisateur en deux classes C_{hot} (pages intéressantes) et C_{cold} (pages sans intérêt), un profil est créé sous forme d'un vecteur de termes sélectionnés de façon à maximiser le gain d'information espéré ($E(w,s)$) de la présence ou l'absence des termes (w) dans l'ensemble des pages s à classifier. Cette fonction correspondant à l'expression :

$$E(w,s) = I(s) - [P(w = present)I(s_{w=present}) + P(w = absent)I(s_{w=absent})]$$

où $I(s) = \sum_{c\in\{C_{hot},C_{cold}\}} -P(s_c)\log_2 P(s_c)$ et $P(w = present)$ est la probabilité que le terme w soit présent dans la page s, et $P(s_c)$ est la probabilité d'obtenir la classe $c \in \{hot, cold\}$ tel que s_c est l'ensemble des pages appartenant à la classe c. Lors de l'analyse, une page est attribuée à l'une des deux classes à l'aide d'un classificateur Bayésien, entraîné à partir d'exemples initialement fournis par l'utilisateur.

Le système **Alipes** [194] cherche de plus à identifier et suivre activement les intérêts de l'utilisateur. Il est particulièrement intéressant à ce titre, en étant parmi les premiers systèmes avec les travaux de [26] à introduire explicitement à la fois une prise en compte des contre-exemples (c'est-à-dire des documents jugés non pertinents par l'utilisateur) et la combinaison de composantes à long terme et à court terme dans la représentation des intérêts de l'utilisateur. Il s'agit d'un système multi agents, gérant un profil utilisateur contenant plusieurs catégories construites par le système, chaque catégorie TDR ayant 3 dimensions correspondant chacune à un descripteur ayant la forme d'un vecteur de poids associés à des termes : un descripteur à long terme V_{LTD}, un descripteur à court terme positif V_{POS} et un descripteur à court terme négatif V_{NEG}. La pertinence d'un document D est évaluée en le comparant aux différents vecteurs représentant les catégories en utilisant l'expression suivante :

$$I_{TDR}(D) = \eta sim(V_{LTD}, V_D) + (1-\eta)(w_{POS} sim(V_{POS}, V_D) - w_{NEG} sim(V_{NEG}, V_D))$$

où sim est la mesure du cosinus entre les deux vecteurs représentant d'une part le descripteur considéré et d'autre part le document, w_{POS} et w_{NEG} sont les poids associés aux descripteurs à court terme positif et négatif, respectivement, et η est un paramètre déterminé expérimentalement, compris entre 0 et 1. La catégorie à laquelle est rattaché le document est choisie en prenant la catégorie maximisant cette pertinence. Le système prend en compte le *feedback*

de l'utilisateur via ses interactions avec les documents proposés pour ajuster les poids associés aux différentes catégories et aux termes dans les catégories. La création d'une nouvelle catégorie est décidée quand la similarité maximum obtenue $I_{TDR}(D)$ est inférieure à un seuil prédéfini, et que le nombre maximum de catégories n'a pas été atteint.

4.2.1.2 Recommandation basée sur la collaboration

Dans cette approche les utilisateurs du système participent activement à l'alimentation d'une base de données gérée par le filtre, contenant des informations sur eux-mêmes, et sur les documents qu'ils ont consulté. Les recommandations pour le nouvel utilisateur sont basées sur ces prédictions. Ainsi, tous les utilisateurs peuvent tirer profit des évaluations des autres en recevant des recommandations pour lesquelles les utilisateurs les plus proches ont émis un jugement de valeur favorable, et cela sans que le système dispose d'un processus d'extraction du contenu des documents.

Ce type de recommandation est suivi par le système ***WebWatcher*** [5]. WebWatcher, assimilable à un guide de musée, se démarque en proposant une fonction d'apprentissage de la qualité des liens connectant les pages entre elles. Le système observe le comportement (le parcours de navigation) de l'utilisateur et lui recommande les liens vers les documents pertinents en fonction de son profil et des profils similaires des autres utilisateurs. Chaque utilisateur définit en quelques termes son objectif, puis navigue en utilisant le système. Ce dernier utilise ces définitions pour étiqueter les liens et attribuer des poids aux termes associés ainsi à un lien (trouvé dans la phrase contenant le lien, ou les titres des chapitres contenant ce lien). Ces annotations sont ensuite utilisées dans le processus de recommandation en comparant les annotations avec les centres d'intérêts initialement exprimés par l'utilisateur. Le système propose à chaque page les trois liens jugés les plus en rapport avec l'objectif énoncé par l'utilisateur.

Une autre approche proposé dans la plateforme **COCoFil** (Community-Oriented Collaborative Filtering) [54] consiste à tirer profit de la présence des utilisateurs dans le système en s'intéressant à leur organisation en communautés. En effet, elle intègre des fonctionnalités destinées à mieux exploiter la notion de communauté d'utilisateurs.

Chaque utilisateur U peut décider des informations qu'il souhaite divulguer ou non, qu'il s'agisse d'informations relatives à son identité, ou même des évaluations qu'il fait pour chaque document. Le système propage ces contraintes de confidentialité (privacy), limitant ainsi certaines fonctionnalités de nature à faire connaître les utilisateurs entre eux. Ces dernières fonctionnalités, particulièrement orientées vers la communauté, permettent notamment de connaître les utilisateurs qui ont attribué une note semblable à U pour un document donné, ou ceux qui d'une manière plus globale ont un profil similaire à celui de l'utilisateur U.

4.2.2 Approches d'appariement personnalisé de l'information

Dans le cadre d'un accès personnalisé, le processus d'appariement classique document-requête est augmenté par les informations contenues dans le profil de l'utilisateur. Le principe de base consiste à définir une fonction d'appariement calculant un score de pertinence document-requête intégrant la composante utilisateur U : $RSV(D,Q) \Rightarrow RSV(D,Q,\mathbf{U})$. Les stratégies de sélection personnalisée dépendent essentiellement du modèle de représentation des composantes informationnelles du modèle d'accès (document, requête, profil utilisateur).

Ce processus d'appariement s'effectue de deux manières :

* Exploiter directement le *contenu* informationnel du profil dans le formalisme de la fonction de pertinence. Il est à noter qu'il existe peu de systèmes qui fournissent réellement une telle approche de personnalisation, néanmoins on trouve des travaux intéressants pour la personnalisation qui exploiter les fondement de la théorie des probabilité. Parmi les approches prédominantes, on trouve les modèles probabilistes d'analyse sémantique latente[1] (PLSA).
* Exploiter la *structure* du contenu informationnel représentant d'une part le profil utilisateur et d'autre part la collection de documents interrogé (essentiellement le *web*). Ceci, en analysant généralement la structure des liens existants dans la topologie du *web* et en les combinant avec celle du profil utilisateur. Il est également à noter qu'il existe peu de systèmes qui fournissent réellement une telle approche de personnalisation, néanmoins des variantes de l'algorithme de *PageRank* [95] offrent des possibilités intéressantes pour la personnalisation.

Nous détaillons dans ce qui suit ces deux directions de recherche en décrivant l'exemple de certains systèmes et en précisant pour chaque système présenté, le type des sources d'évidence apporté par le profil utilisateur dans le processus de sélection de l'information.

4.2.2.1 Approches basées sur le contenu

Dans cette section, nous décrivons les travaux effectués par [111] pour personnaliser la fonction de pertinence. Le modèle d'accès personnalisé proposé se base sur PLSA pour identifier l'intention de recherche de l'utilisateur à partir de son comportement.

PLSA est une extension probabiliste à l'analyse sémantique latente (LSA) [74; 75]. En raison de sa flexibilité, le modèle PLSA a été employé avec succès dans divers domaines d'application, y compris la RI [74]. Mais ce n'est que récemment qu'elle a été appliquée dans des approches de personnalisation de l'accès à l'information [86]. PLSA est un modèle statistique de mélange de classe d'objet latente. L'hypothèse de base est qu'il existe un ensemble de facteurs cachés dans les cooccurrences des deux ensembles d'objets. Cela signifie que les occurrences des deux ensembles d'objets sont indépendantes quand les variables latentes sont données.

[1] Probability Latent Semantic Analysis

Dans ces travaux, le profil est construit implicitement. Il contient les documents sélectionnés (par click, nommé *cliqués* documents) pour chaque requête, et la requête soumise. Ainsi, l'utilisateur, les requêtes et les *clicked* documents sont collectés comme un triplet d'informations cooccurrentes dans le profil. Où l'utilisateur $u \in U\{u_1, u_2, ..., u_n\}$, la requête $q \in Q\{q_1, q_2, ..., q_m\}$ et les pages associés $p \in P\{p_1, p_2, ..., p_l\}$. Les relations sont associées aux variables latentes $z \in Z\{z_1, z_2, ..., z_k\}$.

Le traitement de ces données s'effectue de deux manières. Si la requête de l'utilisateur a déjà été soumise, le système est en mesure de retrouver dans le profil les documents pertinents les plus fréquemment visités par l'utilisateur. Par contre, si c'est une nouvelle requête, le problème qui se pose est : *Etant donné les documents déjà sélectionnés, quelle est la pertinence des documents qui devraient être sélectionnés en réponse à cette requête*.

Formellement, le problème peut être décrit comme suit : étant donné un ensemble T contenant les données collectées représentées par des triplets (u, q, p), une fonction de mapping $f : U * Q \rightarrow P$ doit être apprise par le modèle. La résolution de ce problème est donnée en appliquant le modèle de PLSA. Il est à noter que lors d'un scénario de recherche sur le *web*, une observation, au sens PLSA, est un triplet (u, q, p) correspondant à l'événement qu'un utilisateur u soumet une requête q à un moteur de recherche , et sélectionne une page p à partir des résultats retournés (MSN Search, dans ce cas). La fonction f permet de sélectionner pour chaque paire d'information "*utilisateur, requête*" notée (u, q) contenue dans le profil, une nouvelle page p' n'ayant pas été déjà collectée dans le profil (c-à-d, que le triplet $(u, q, p') \notin T$), tel que la page p' sélectionnée est celle ayant une forte probabilité de pertinence.

Ce modèle d'appariement dépend des hypothèses d'indépendance conditionnelle entre ces éléments, à savoir que chaque ensemble d'objets observés est conditionné par l'état des variables latentes associées. En tenant compte de ces hypothèses, les requêtes et les pages *web* sont indépendantes quand les intentions de recherche sont observées. Cela signifie qu'un utilisateur u et une requête q déterminent une intention de recherche latente z. Cette variable latente sert à sélectionner la page *web* p.

A partir de la topologie du modèle, représentée par la figure 4.2, la probabilité jointe du modèle est obtenue donnée par a fonction suivante :

$$P(u, q, p) = \sum_{z \in Z} P(z)P(u|z)P(q|z)P(p|z)$$

L'estimation des probabilités attachées à chaque membre de cette formule est obtenue en appliquant l'algorithme *Expectation-Maximization* (*EM*) [52]. L'algorithme alterne deux étapes :

- l'étape d'expectation (E) : où des probabilités *a posteriori* sont calculées pour la variable latente sur la base des évaluations courantes des paramètres,
- l'étape de maximisation (M), où les paramètres sont réestimés pour maximiser l'expectation des probabilités globales des données.

Ainsi, le modèle peut calculer la probabilité de pertinence d'une page p sachant l'utilisateur

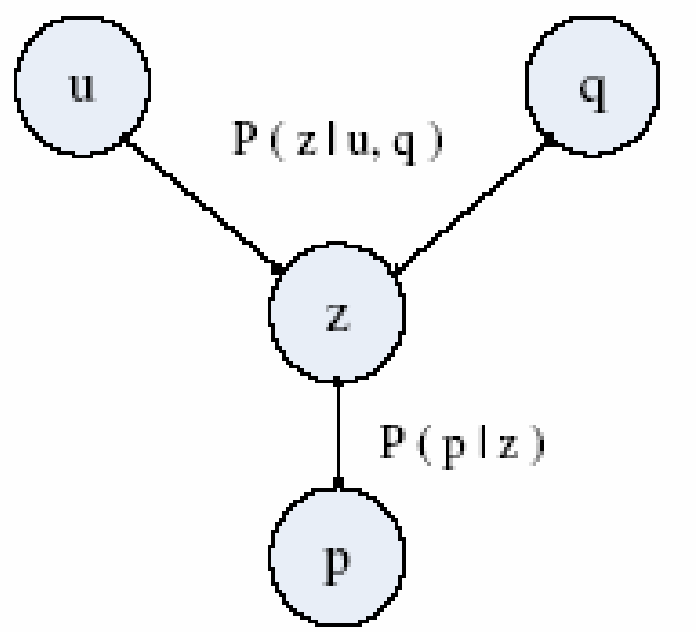


Figure 4.2 — Représentation graphique du modèle pour un triplet (u,q,p) [111]

u et la requête q sur la base de la formule suivante :

$$P(p|u,q) = \frac{\sum_{z\in Z} n(u,q,p)P(z|u,q,p)}{\sum_{p'}\sum_{z\in Z} n(u,q,p')P(z|u,q,p')} \tag{4.1}$$

$n(u,q,p)$ est le nombre de fois où l'utilisateur u sélectionne la page p pour la requête q. Puis, le système ordonne les résultats sur la base des valeurs de ces probabilités et retourne les pages ayant les valeurs les plus élevées.

4.2.2.2 Approches basées sur la structure

Le processus de RI traditionnel se base seulement sur le contenu des documents pour décider de la pertinence des résultats en réponse à une requête. Dans un environnement hypertexte tel que le *web*, une métrique additionnelle peut être introduite basée sur la structure des liens entre les pages. Parmi les plus connues, il y a les mesures *Hub* et *Autority* proposées par [95] (également nommé : *Hyperlinked Induced Topic Search* (HITS)), et aussi le *PageRank* qui constitue un composant important dans l'algorithme de recherche utilisé par Google [31].

Plus précisément, le *PageRank* (PR) est un vote assigné à une page A collectée de toutes les pages *web* $T_1,..,T_n$ qui se dirigent vers la page A. Il représente l'importance de la page dirigée.

Le PageRank de la page A est calculé par la formule suivante :

$$PR(A) = (1-d) + d[\frac{PR(T_1)}{C(T_1)} + ... + \frac{PR(T_n)}{C(T_n)}]$$

où le paramètre d est le facteur d'atténuation prenant ses valeurs entre 0 et 1 et $C(T_n)$ est défini comme le nombre de liens dans la page T_n.

Les valeurs de PR fournissent des évaluations *a priori* d'importance pour toutes les pages *web*, indépendant de la requête. Lors de la recherche, ces valeurs d'importance sont combinées avec les degrés de pertinence obtenus par une RI traditionnelle pour ordonner la pertinence des résultats.

L'algorithme du *topic-sensitive PageRank* (TSPR) proposé par [72] est une extension intéressante du PageRank, qui peut potentiellement fournir différents rangs pour différentes requêtes pour chaque utilisateur, tout en maintenant l'avantage de l'efficacité du PageRank standard. Dans TSPR les pages considérées importantes pour certains domaines ne peuvent être considérées importantes pour d'autres selon chaque utilisateur. Pour cette raison, l'algorithme calcule 16 ensembles de valeurs de domaines sensibles de PageRank, chacune basée sur les URLs des catégories de haut niveau d'ODP. Lors de la recherche, chaque requête soumise est associée au domaine le plus approprié et, au lieu d'employer une valeur globale simple de PageRank, une combinaison linéaire des rangs des domaines est calculée et pondérée en utilisant les similitudes de la requête avec les domaines. Formellement, pour calculer le TSPR concernant le domaine t, on définit en premier un vecteur aléatoire de probabilité du passage de la page v au domaine t, $\mathbf{E}_t = [E_t(1), ..., E_t(n)]$, comme ceci :

$$E_t(v) = \begin{cases} 1/v_t \textit{ si page v est relie au domaine t} \\ 0 \textit{ sinon} \end{cases} \tag{4.2}$$

Où n_t est le nombre total de pages reliées au domaine t. L'ensemble de pages considérées connexes au domaine t (c'est à dire, les pages dont la valeur du $E_t(v) \neq 0$) sont qualifiés par l'ensemble polarisé du domaine t. Ainsi, le score de TSPR de la page v en considérant t est défini par :

$$TSPR_t(v) = d * \sum_{v \in A_v} TSPR_t(v)/l_v + (1-d) * E_t(v) \tag{4.3}$$

Où v est la page concernée, A_v l'ensemble des pages liées à v, l_v le nombre de liens sortant de la page v, d la probabilité que l'utilisateur suit le lien sortant et $1-d$ la probabilité restante du saut de page.

Etant donné que tous les calculs basés sur les liens s'effectuent offline, le temps d'exécution de ce processus de recherche est comparable à l'algorithme original de PageRank. Les expérimentations menées sur ce système ont conclu que l'utilisation de domaines-spécifiques de PageRank peut améliorer l'exactitude de recherche sur le *web*.

Une variante de l'approche du topic-sensitive PageRank est présentée par [84]. L'algorithme qu'il propose (PPV), adapte le principe de calcul de l'*Authority* d'une page, donné par l'algorithme *PageRank*, en utilisant une distribution de probabilité tenant compte de la présence de liens entrants et de liens sortants vers et/ou les pages favorites de l'utilisateur. Plus précisément, pour un utilisateur u donné, l'équation posée par l'algorithme PPV est : $v = (1-c)Av + cu$ où c est une constante avec $c \in (0,1)$, A est la matrice de contingence du graphe et u est un vecteur de préférences de l'utilisateur où chaque composante traduit le poids de préférence de l'utilisateur pour chacune de ses pages favorites.

Les vecteurs personnalisés de PageRank (PPVs) ne sont pas produits à partir de zéro à chaque fois que l'utilisateur soumet une requête, dans le but de diminuer la complexité des calculs. En effet, un ensemble partiel de vecteurs de PageRank représentant les pages centrales est initialement calculé à partir de la topologie du *web*, puis les PPVs sont calculés en appliquant une somme linéaire de ces vecteurs de base. De ce fait, on évite l'étape d'analyse de lien nécessaire pour la génération des PPVs au moment de l'exécution de chaque requête.

Basé sur l'algorithme PPV, [41] proposent une plateforme de personnalisation de score appelée PROS qui fournit un score personnalisé aux pages *web* selon les profils utilisateur construits automatiquement, à partir des annotations de l'utilisateur ou l'ensemble des pages *web* fréquemment visitées. En d'autres termes, dans PROS, les pages jugées les plus intéressantes pour l'utilisateur sont transmises à un module nommé *HubFinder*. Son rôle est de rassembler des pages centrales (Hub) liées aux domaines du profil de l'utilisateur (les pages contenant de nombreux liens à des ressources de qualité). Ce module analyse uniquement la structure des liens du *web* en appliquant une version adaptée aux besoins de l'utilisateur de l'algorithme HITS proposé par [95].

4.2.3 Modèle de ré-ordonnancement des résultats de recherche

Un SRI de base retourne à l'utilisateur une liste de documents ordonnés selon leur degré de pertinence en réponse à sa requête. Cette requête constitue la principale source d'évidence pour déterminer la pertinence des documents. Le problème posé est que l'utilisateur n'ayant aucune (voire peu de) connaissance sur la collection de documents et de l'environnement de recherche, il lui est difficile de formuler des requêtes claires et appropriées, permettant de cibler la recherche uniquement aux documents pertinents [147]. Par conséquent, les documents pertinents se trouvent mélangés, lors de la présentation des résultats, aux documents non pertinents. En effet, on estime que parmi les vingt principaux résultats des SRI traditionnels, seulement la moitié d'entre eux est réellement en adéquation avec les besoins de l'utilisateur. En outre, l'ordonnancement des résultats ne dépend pas exclusivement d'une mesure globale de pertinence, où le rang identiquement calculé pour l'ensemble des utilisateurs est supposé approprié quelque soit l'utilisateur.

La personnalisation à ce stade du processus de recherche offre une solution en réordonnant les résultats pour ne présenter à l'utilisateur que les documents pertinents en réponse à son besoin en information. Ce besoin est formulé en conjuguant les informations données par

l'utilisateur durant la session de recherche tel que les requêtes soumises et celles extraites de son profil représentant ses besoins récurrents (historique, centres d'intérêts, etc.). Ainsi, la restitution des résultats s'effectue en fonction de la notion de pertinence personnelle de l'utilisateur où le rang du document est calculé en corrélation avec un utilisateur spécifique sur la base de son contexte d'interaction. Ainsi, l'idée principale du ré-ordonnancement est d'intégrer une mesure de corrélation entre le profil utilisateur et chaque document comme facteur de distinction dans le calcul du rang.

Parmi les travaux les plus représentatifs, nous présentons ceux menés par [166]. Ils proposent un système d'accès personnalisé nommé " *mysearch* " qui réordonne les résultats de recherche obtenus par un moteur de recherche externe, en l'occurrence Google. Le système attribue le rang le plus élevé aux documents correspondant aux centres d'intérêts du profil de l'utilisateur. Le profil exploité dans ce système est basé sur une représentation conceptuelle (hiérarchie de concepts pondérés), crée à partir d'ODP et l'historique de recherche de l'utilisateur.

Lors de la soumission d'une requête, les 10 meilleurs résultats de recherche sont classifiés, sur la base d'une fonction de similarité, dans la même hiérarchie de concept de référence que le profil de l'utilisateur. Les titres et les résumés de ces résultats sont classés pour créer un profil de document dans le même format que le profil utilisateur. Le profil document est ensuite comparé au profil utilisateur pour calculer le rang conceptuel entre chaque document et les centres d'intérêts de l'utilisateur.

Le poids final attribué au document utilisé pour le réordonnancement (de sorte que les résultats assimilés aux meilleurs centres d'intérêt du profil soient rangés en haut de la liste) est calculé en combinant le rang conceptuel obtenu avec le rang initialement donné par Google, en utilisant la fonction de pondération suivante :

$$RangFinal(u_i,d_j) = \alpha * RangConceptuel(u_i,d_j) + (1-\alpha) * RangInitial(u_i,d_j) \qquad (4.4)$$

Tel que $RangConceptuel(u_i,d_j)$ est le rang du document d_j obtenu en calculant une similarité entre le profil document et les concepts du profil utilisateur u_i, selon la formule de similarité suivante : $Sim(u_i,d_j) = \sum_{k=1}^{N} Wt(i,k) + Wt(j,k)$, où N le nombre totale des concepts du profil utilisateur i, $Wt(i,k)$ est le poids du concept k dans le profil utilisateur i et $Wt(j,k)$ est le poids du concept k dans le profil document j ; α une valeur constante entre 0 et 1 ; et $RangInitial(u_i,d_j)$ est le rang initialement attribué au document par le moteur de recherche.

Lorsque α égale 0, le rang conceptuel ne donne aucun poids et la valeur d'appariement est équivalente au rang original affecté le moteur de recherche. Si α a une valeur égale à 1, le rang initial est ignoré et on considère uniquement le rang conceptuel. Évidemment, les deux rangs, conceptuel et attribué par le moteur de recherche, peuvent être combinés selon différentes proportions en changeant la valeur de α.

4.2.4 Modèle de la reformulation de requêtes

Les requêtes utilisateur sont assurément une source évidente importante pour l'identification des besoins en information de l'utilisateur. Néanmoins, comme nous l'avons déjà mentionné, les utilisateurs soumettent souvent des requêtes très courtes et ambiguës. L'objectif de la personnalisation à ce stade du cycle de vie de la requête est de clarifier le besoin en information de l'utilisateur en se basant sur ce que le système a appris à son sujet. Ainsi, la reformulation de requête dans ce cadre intègre les composantes informationnelles issues du profil de l'utilisateur pour identifier, enrichir et cibler son intention de recherche.

Dans le cadre d'un accès personnalisé aux bases de données, l'approche de reformulation de requête consiste à proposer des algorithmes d'enrichissement de la requête utilisateur par des prédicats de préférence candidats choisis du profil. Ces prédicats sont souvent affectés d'un poids traduisant son degré d'importance dans le profil [104]. La reformulation peut également être réalisée en appliquant des algorithmes de réécriture pour traduire la requête en un ensemble d'expressions pouvant être évaluées sur les sources de données. Récemment, [103] propose une nouvelle approche de *RR* alternant des étapes d'enrichissement et de réécriture afin de mieux tenir compte du profil utilisateur. Elle se compose de quatre étapes : expansion de la requête, identification des sources pertinentes, combinaison des sources pertinentes et enrichissement final. On trouve justement dans ces travaux une synthèse des approches de reformulation de requête pour l'accès personnalisé aux bases de données.

D'un point de vue RI, l'approche de personnalisation suivie consiste à exploiter l'historique de recherche de l'utilisateur pour clarifier le but de recherche d'utilisateur. Dans l'approche d'expansion de requête personnalisée proposée par [51] le profil, construit à partir de l'ensemble des fichiers *logs* issus des interactions de l'utilisateur avec le SRI, est exploiter pour identifier des corrélations entre les termes des requêtes de l'utilisateur. Ces corrélations sont établies à travers les fichiers *logs* entre les termes de la requête et ceux des documents de la collection. Ainsi, cette approche peut être vue comme une technique de construction de thésaurus intermédiaire entre les espaces vectoriels engendrés par les termes des documents et la requête.

Cette approche se base sur le principe que, si un ensemble de documents est souvent sélectionné pour une même requête, alors les termes de ces documents sont fortement liés aux termes contenus dans la requête. Ainsi, les relations sont extraites pour déduire la distance entre les espaces de la requête et les documents. Pour cela, ils proposent d'exploiter les fichiers logs contenus dans le profil de l'utilisateur pour établir ce pont entre les deux espaces. Ce processus débute par l'extraction d'une session de recherche pour chaque requête à partir de l'ensemble de tous les fichiers logs collectés. Chaque session est identifiée comme suit :

session := < texte de requête > [document sélectionné]*

La session contient une requête ainsi que l'ensemble des documents jugés pertinents que l'utilisateur a sélectionnés (en cliquant dessus). Comme illustré par la figure 4.3, des liens pondérés peuvent être créés entre l'espace de requête (tous les termes de la requête) et les

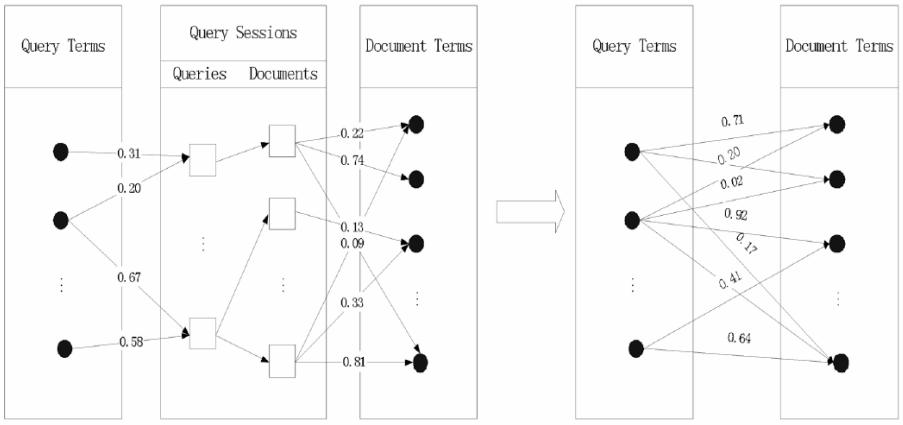


Figure 4.3 — Corrélations établies entre les termes de la requête et du document via les sessions de requêtes

sessions de requête et entre l'espace de document (tous les termes du document) et les sessions.

Le processus d'expansion de la nouvelle requête Q s'effectue selon les étapes suivantes :

1. Extraction de tous les termes de la requête Q ;
2. Identification des documents correspondant à chaque terme de la requête dans les sessions de requête ;
3. Pour chaque terme dans ces documents, appliquer la formule suivante pour évaluer le degré d'évidence inhérent à la sélection de ce document pour l'expansion de cette requête. Cette fonction mesure un poids de cohésion en combinant les relations de cooccurrences entre le terme et toute la requête :

$$CoWeight_Q(w_j^{(d)}) = \ln(\prod_{w_t^{(q)} \in Q} (P(w_j^{(d)}|w_t^{(q)}) + 1)) \tag{4.5}$$

   Où $P(w_j^{(d)}|w_t^{(q)})$ est la probabilité conditionnelle mesurant le degré de corrélation entre chaque terme du document $w_j^{(d)}$ et chaque terme requête $w_t^{(q)}$;
4. Sélectionner n termes de l'espace document ayant le plus fort poids de cohésion et formuler une nouvelle requête Q' en ajoutant ces termes à la requête Q ;
5. Soumettre cette requête Q' au SRI pour lancer la recherche.

En 2004, [190] proposent également une approche d'expansion de requête personnalisée en se basant sur les probabilités. Comparativement à l'approche précédente, [190] augmentent la représentation de la session de recherche avec la notion de contexte utilisateur. Ainsi, une session de recherche extraite des les fichiers logs des requêtes soumises par l'utilisateur est représentée par une séquence requête-contexte-document sous la forme suivante :

Session requête := <requête, contexte> [documents cliqués]*

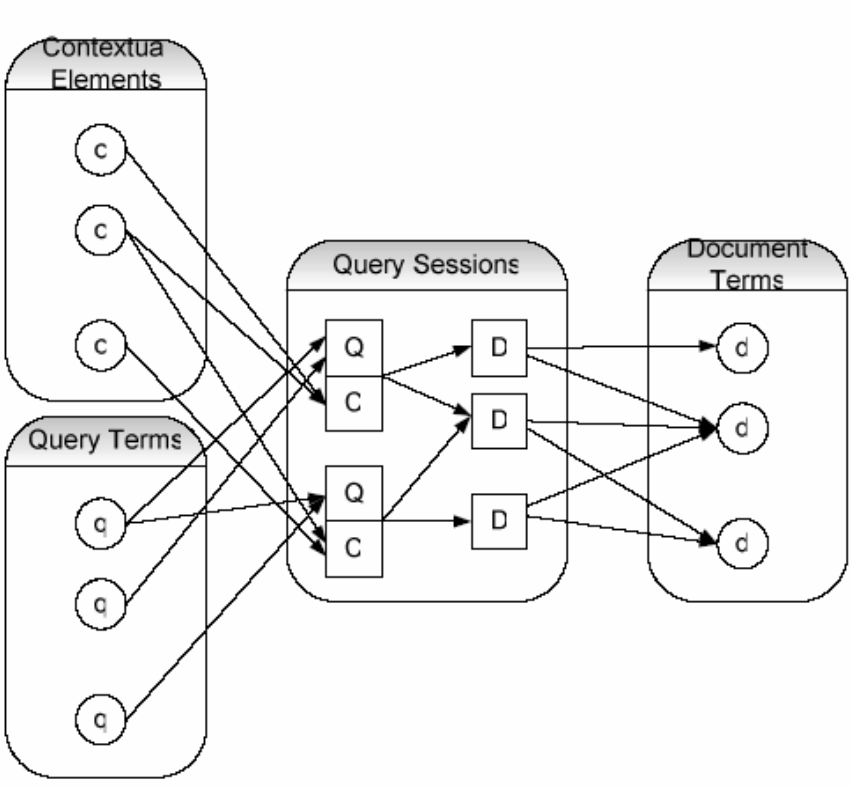


Figure 4.4 — Exemple d'une session de recherche

Les documents *cliqués* correspondent à l'ensemble de documents que l'utilisateur a sélectionnés ou annotés lors de différentes sessions de recherche.

L'idée centrale de la méthode est que si un ensemble de documents est souvent choisi pour les requêtes similaires et lors de contextes d'utilisation semblables, les termes de ces documents sont fortement liés aux termes formulant les requêtes et les éléments du contexte. En outre, si les requêtes similaires et les contextes semblables co-occurrents fréquemment dans les notations de l'utilisateur, les termes des requêtes sont en étroite corrélation avec les éléments contextuels. Ainsi, des corrélations probabilistes parmi cet ensemble informationnel peuvent être établies sur la base des notations, tel que cela est illustré par la figure 4.4. La figure représente une session de recherche établissent les corrélations entre les termes de la requête, les éléments contextuels et les termes d'un document.

L'information mutuelle est utilisée pour déterminer les degrés de corrélation entre les termes de la requête, les éléments contextuels et les termes des documents (Figure 4.5). L'information mutuelle est une mesure de dépendance statistique entre deux variables aléatoires basée sur l'entropie de Shannon. Il s'agit d'un score d'association de deux mots (x,y) noté IM qui permet de comparer la probabilité d'observer ces deux mots ensemble avec la probabilité de les observer séparément. Selon [42], la définition du score IM est la suivante : $IM(x,y) = log_2(P(x,y)/(P(x)P(y)))$ où $P(x)$ et $P(y)$ sont les probabilités d'observer les mots x et y, et $P(x,y)$ est la probabilité de les observer simultanément.

Si $IM(x,y)$ est fortement positive, cela signifie que x et y apparaissent très souvent ensemble. Si $IM(x,y)$ est proche de 0, alors x et y n'ont aucun rapport et enfin, si $IM(x,y)$ est fortement négative, alors x et y ont des distributions complémentaires.

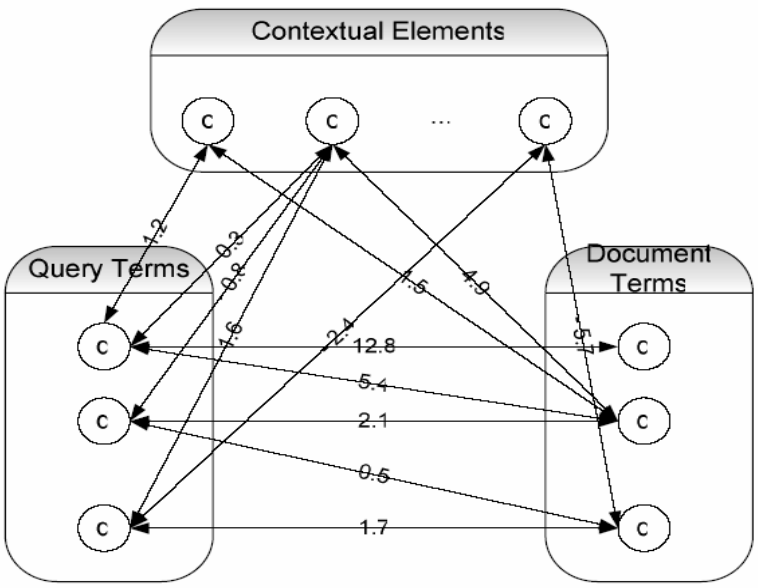


Figure 4.5 — Information mutuelle entre les termes de la requête, les éléments contextuels et les termes des documents

Dans la suite et sur la base de ces corrélations, quatre modèles ont été proposés pour générer des expansions de termes de la requête : un *modèle orienté contexte*, un *modèle indépendant requête-contexte*, un *modèle dépendant requête-contexte* et un *modèle filtrant le contexte*

– *Modèle 1 : modèle orienté contexte*

Le premier modèle est assez simple et intuitif : les termes des documents bien corrélés au contexte sont choisis comme expansion de termes pour modifier la requête initiale :

$$M_1(d \triangleleft Q, C >) = I(d,C) = \sum_i I(d,c_i)$$

où $I(d,C)$ est l'information mutuelle entre le contexte et les termes du document. Dans ce modèle les termes utilisés pour l'expansion de requête ne proviennent pas directement du contexte mais des documents corrélés au contexte.

– *Modèle 2 : modèle indépendant requête-contexte*

Le principal défaut du premier modèle est que les termes de l'expansion sont générés uniquement à partir du contexte et ne sont donc pas corrélés à la requête. Le deuxième modèle utilise la requête et le contexte pour contrôler le processus d'expansion de requête.

$$M_2(d \triangleleft Q, C >) = I(d, < Q, C >)$$

$$M_2(d \triangleleft Q, C >) = I(d,C) + I(d,Q)$$

$$= \sum_i I(d,c_i) + \sum_i I(d,q_i)$$

où $I(d,Q)$ est l'information mutuelle entre la requête et les termes du document.
L'avantage du deuxième modèle par rapport au premier réside dans le fait qu'il affecte des poids plus élevés aux termes des documents qui sont corrélés à la fois à la requête et au contexte.

– *Modèle 3 : modèle dépendant requête-contexte*
Pour diminuer l'effet de l'hypothèse d'indépendance entre la requête et le contexte, un troisième modèle a été introduit pour prendre en compte les relations de dépendance entre la requête et le contexte. Ce modèle est défini par :

$$M_3(d \triangleleft Q, C >) =$$

$$= I(d, < Q, C >) = \sum_i I(d,c_i) + \sum_j I(d,q_j) + \sum_{ij} I(d, < q_j, c_i >)$$

où $\sum_{ij} I(d, < q_j, c_i >)$ est l'information mutuelle entre un terme de document et une paire requête-contexte. C'est ce facteur qui introduit une dépendance requête-contexte. Le paramètre est introduit pour ajuster le poids de ce facteur de dépendance requête-contexte.

– *Modèle 4 : modèle filtrant le Contexte*
Un problème commun entre les trois premiers modèles est que le bruit dans un contexte n'est pas traité. Manifestement, le bruit du contexte peut facilement produire des expansions de termes hors de propos. Ainsi, le quatrième modèle utilise l'information mutuelle entre requête et contexte pour éliminer le bruit dans les éléments du contexte.
$M_4(d \triangleleft Q, C >) = I(d, < Q, C >) = I(d, < Q, C' >) = \sum_i I(d,c_i) + \sum_j I(d,q_j) + \sum_{ij} I(d, < q_j, c_i >)$
où $C' = \{c/c \in C, I(c,Q) \geq \alpha\}$ $I(c,Q)$ est l'information mutuelle entre la requête et un élément contextuel. Elle est utilisée pour filtrer des éléments contextuels non corrélés à la requête courante, α représente le seuil pour le processus de filtrage.

Les corrélations ainsi construites vont ensuite être utilisées dans le modèle d'accès pour estimer la "*bonne*" pertinence d'un document pour une requête et un contexte donné. En effet, l'approche d'expansion de requête est basée sur le calcul des probabilités décrites ci-dessus. Lorsqu'une nouvelle requête est soumise dans le contexte, une liste des termes des documents corrélés est sélectionnée, puis ordonnée selon les probabilités conditionnelles obtenues à travers le modèle. Enfin, les termes ayant obtenu le plus haut rang sont alors utilisés pour reformuler la requête de l'utilisateur.

4.3 Evaluation des systèmes d'accès personnalisé à l'information

L'évaluation des SRI est depuis le début des travaux sur la RI un des piliers de l'évolution de ce domaine. La qualité de l'évaluation est d'une importance capitale puisqu'elle permet de discriminer les différents modèles. La démarche de validation en RI se base sur l'évaluation expérimentale des performances du modèle ou du système proposé, selon le modèle de Cranfield [45]. L'évaluation des performances d'un modèle de RI, permet de paramétrer le modèle, d'estimer l'impact de chacune de ses caractéristiques. Dans la plupart des protocoles d'évaluation des systèmes d'accès à l'information, les mesures sont construites à partir des jugements exprimés par des utilisateurs ou par des experts. Ces mesures concernent plusieurs critères : le temps de réponse, la pertinence, la qualité et la présentation des résultats.

L'évaluation orientée vers l'utilisateur est une composante primordiale dans le cadre de l'accès personnalisé à l'information. En effet, les objectifs d'une telle évaluation sont de mesurer l'adéquation des profils utilisateur construits par le système avec les centres d'intérêts effectifs de l'utilisateur ; ainsi que l'impact de l'intégration de ce profil, dans le processus d'accès, sur les performances de recherche. Néanmoins, force est de constater que de telles méthodologies d'évaluation restent encore peu formalisées pour être appliquées dans des campagnes d'évaluation effectives.

Nous présentons dans cette section une synthèse des approches d'évaluation utilisées dans le cadre de l'accès personnalisé. Nous décrivons en premier lieu, le protocole d'évaluation standard TREC (Text Retrieval Conference) dédié à la RI traditionnelle. En second lieu, nous dressons un bilan des limites du protocole TREC [2] à travers la problématique liée à la mise en place de la campagne d'évaluation standard et formelle pour l'accès personnalisé. Puis nous présentons les éléments communs des approches d'évaluation utilisées dans les travaux de référence dans ce domaine, selon une organisation qui se veut représentative d'un protocole d'évaluation de systèmes d'accès personnalisé à l'information. Nous finirons ensuite par un aperçu de quelques travaux de référence.

4.3.1 Le programme d'évaluation TREC

Des campagnes d'évaluation ont été mises en place au niveau mondial pour offrir un cadre standardisé et formel destiné à des protocoles d'évaluation communs. L'initiative la plus importante actuellement pour la construction de collections de tests est sans conteste TREC.

TREC est un projet international initié au début des années 90 par le NIST [3] aux Etats-Unis dans le but de proposer des moyens homogènes d'évaluation de systèmes documentaires sur des bases de documents conséquentes.

[2] Text Retrieval Conference, *http ://trec.nist.gov*

[3] National Institute of Standards and Technology

Il est co-sponsorisé par le NIST et DARPA/ITO [4]. L'objectif de TREC est d'encourager les travaux de recherche d'information permettant l'accès à des bases volumineuses en fournissant :

* Une base importante de test,
* des procédures d'évaluation uniformes,
* un forum pour les organismes intéressés par une comparaison de leurs résultats.

4.3.1.1 Description d'une tâche TREC

Un ensemble de tâches différentes est proposé aux participants qui soumettent des résultats à autant de tâches qu'ils le souhaitent. Le principe général d'une tâche est que l'on dispose d'une collection de requêtes (ou plus exactement d'expressions de besoins d'information, sans préjuger de la forme que peut prendre la requête effective devant sélectionner les documents), d'une collection de documents et d'un ensemble complet de valeurs de pertinence : toute association requête-document a été jugée soit satisfaisante, soit invalide (selon l'appréciation d'un arbitre ou des assessors).

La tâche *Ad-hoc* dans TREC évalue les performances des SRIs sur des ensembles statiques de documents, seules les requêtes changent. Cette tâche est similaire à une recherche dans une bibliothèque par exemple, où la collection est connue mais les requêtes susceptibles d'être posées ne le sont pas. La tâche (*Ad-hoc*) consiste d'abord à créer des requêtes à partir des besoins en information (*Topics*) posés par de vrais utilisateurs (*assessors*), environ une cinquantaine. Chaque participant fournit au NIST pour l'évaluation la liste des 1000 premiers documents retrouvés par leur système en réponse à chacune de ces requêtes. Les *assessors* jugent la pertinence des 100 à 200 premiers documents de chaque système puis différentes mesures d'évaluation sont calculées (le rappel et précision, la précision moyenne, la précision à 10, 20, 30 etc.).

4.3.1.2 Collections de test

Les collections TREC sont de l'ordre de quelques giga-octets et de quelques centaines de giga-octets pour les VLC (Very Large Collections et TB Terabyte). Les documents sont issus de différentes sources dont essentiellement la presse écrite tel que le Wall Street Journal mais également des documents *web*. Ces données sont disponibles sur le serveur du NIST.[5]

1. **Les documents**

   Le corpus a été rassemblé avec un souci de représentativité de la variété des documents rencontrés dans la réalité.

[4] Defense Advanced Research Projects Agency - Information Technology Office

[5] *http ://trec.nist.gov*

Les documents de cette collection proviennent de différentes sources de données : des articles de presse, des résumés courts de publications, des brevets, ainsi que (dernièrement) des documents informatiques mis sur Internet. Il semble que certains soient (faiblement) structurés : présence d'un titre, indication des paragraphes, les autres documents ont des structures hétérogènes annotées de métadonnées. Il existe quatre dimensions de variation :

(a) *longueur* : la très grande majorité des documents (plus de 99 % d'entre eux) sont de l'ordre de 300 mots ou moins : c'est relativement court. Les quelques documents plus longs sont des brevets d'environ 3 000 mots.

(b) *genre* : une petite dizaine de sources sont distinguées ; mais une bonne moitié d'entre elles fournissent des articles de presse. Les autres genres concernés sont des résumés courts de publications, et (marginalement) une collection de documents légaux ou des brevets.

(c) *langue* et *format* : les documents sont essentiellement en anglais, souvent sous le format SGML avec des DTD, ou sous le format Html

(d) *date* : les plus anciens datent de 1987.

```
<DOC>
<DOCNO> AP891231-0001 </DOCNO>
<FILEID>AP-NR-12-31-89 2359EDT</FILEID>
<FIRST> PM-MonkeyBusiness     12-31 0269</FIRST>
<SECOND>PM-Monkey Business,0276</SECOND>
<HEAD>Yacht That Took Gary Hart On Famous Cruise Suffered From Fame</HEAD>
<DATELINE>DENVER (AP) </DATELINE>
<TEXT>
<TEXT>
  Monkey Business, the yacht that helped sink Gary Hart's presidential aspirations in 1988, is for sale, and its
captain says notoriety from Hart's trip to Bimini with Donna Rice hurt business.
  …
</TEXT> ... </TEXT>
</DOC>
```

Figure 4.6 — Un exemple d'un document de collection AP

La figure 4.6 donne un exemple de la structure d'un document issu des articles de presse des AP [6] Newswire, collectés par AT & T Bell Laboratories pour les années 1988 et 1989.

2. **Topics (sujets)**

   Les topics sont des textes à partir desquels les requêtes sont construites. Les topics suivent le modèle de base de TREC illustré par l'exemple suivant :

[6] Associeted Press

```
<top>
<head> Tipster Topic Description
<num> Number: 062
<dom> Domain:  Military
<title> Topic:  Military Coups D'etat
<desc> Description: Document will report a military coup d'etat,
either attempted or successful, in any country.
<smry> Summary: Document will report a military coup d'etat,
either attempted or successful, in any country.
</top>
```

Elles sont définies par :

- **Un titre** : <title >*Topic : Design of the "Star Wars" Anti-missile Defense System* ;
- **Un numéro de requête** : <num> *Number : 101* ;
- et une **description** qui détaille le titre et une partie narrative qui précise exactement les documents qui doivent être pertinents et également ce qui ne doivent pas l'être.

3. **Les jugements de pertinence**

   L'évaluation est réalisée à partir d'un ensemble de documents, d'un ensemble de requêtes et d'un ensemble de jugements (liste des documents pertinents pour une requête donnée). La pertinence d'un document pour une requête est codée par une valeur numérique sur une échelle allant de non pertinent (valeur de 0) à très pertinent (valeur de $+2$). Ces jugements sont regroupés dans des fichiers *Qrels*, dont la structure est la suivante :

   TOPIC ITERATION DOCUMENT# RELEVANCY

   Où,
   - TOPIC : est le numéro de la requête ;
   - ITERATION : est le nombre d'itérations (presque toujours à zéro et non utilisé) ;
   - DOCUMENT# : est le numéro officiel du document, qui correspond au champ «docno » dans les documents ;
   - RELEVANCY : est un code binaire : 0 pour «non pertinent » et 1 pour «pertinent ».

4.3.1.3 Le protocole d'évaluation

Dans la plupart des protocoles, les mesures permettant l'évaluation des SRIs sont construites à partir des jugements de valeurs exprimés par des utilisateurs ou par des experts. Pour une requête et un ensemble de documents proposés en résultats, nous pouvons mesurer les taux de performance des SRI par différentes mesures d'évaluation. Les mesures de *précision* et *rappel* sont les deux métriques les plus utilisées en RI.

1. **Mesures de précision et rappel** Les mesures de précision/rappel sont obtenues en partitionnant l'ensemble des documents, restitués par le SRI, en deux catégories : les documents pertinents et les documents non pertinents. Ces deux catégories se définissent comme suit :
   - **Taux de précision** : le taux de précision mesure la capacité du système à rejeter tous les documents non pertinents à une requête. Il est donné par le rapport entre l'ensemble des documents sélectionnés pertinents P_r et l'ensemble des documents sélectionnés D_r.
   $$precision = \frac{P_r}{D_r} \tag{4.6}$$
   - **Taux de rappel** : le taux de rappel mesure la capacité du système à retrouver tous les documents pertinents répondant à une requête. Il est donné par le rapport entre les documents retrouvés pertinents P_r et l'ensemble des documents pertinents de la collection R.

   $$rappel = \frac{P_r}{R} \tag{4.7}$$
2. **Mesures à X documents et la précision moyenne** Deux mesures, communément utilisées dans le cadre de TREC, sont la précision à X documents noté PX (X peut prendre différentes valeurs : 5, 10, 15, ... ,1000) et la précision moyenne (ou *MAP* pour Median Average Precision).
   La précision à X documents représente le nombre de documents pertinents sur les X premiers. Elle est souvent reliée à ce que l'on appelle la *précision exacte* ou la *R-precision*. La précision exacte représente celle obtenue à l'endroit où elle vaut le rappel. Si la requête admet n documents pertinents, la précision exacte est celle calculée pour les n premiers documents de la liste ordonnée des documents restitués.
   La précision moyenne est la moyenne des valeurs de précision à chaque document pertinent de la liste ordonnée. Elle tient compte à la fois de la précision et du rappel. Elle représente la moyenne des précisions calculées pour chaque document pertinent à trouver, au rang de ce document.

Les résultats de l'évaluation des performances du SRI sont obtenus en comparant les requêtes fournis par le système relativement à celles attendues et ce, en utilisant les mesures ci-dessus. Ainsi, chaque système restitue pour chaque requête 1000 documents, classés par ordre de pertinence. On examine alors comment varient les indicateurs ci-dessus quand on considère les 1, 2, ..., ou les 1000 premiers documents. La qualité des résultats d'un système est alors représentée par la courbe de rappel en fonction de la précision.

4.3.2 Problématique de l'évaluation d'une tâche d'accès personnalisé

Le domaine de la RI a une tradition bien établie d'évaluation expérimentale, qui remonte aux expérimentations de Cranfield, et qui continue à travers les compagnes TREC. Comme

nous venons de le présenter l'approche générale d'évaluation de la recherche *adhoc* s'effectue sur des collections statiques pour retourner les documents pertinents pour une requête (topic) préalablement connue. Elle exige trois composants : une collection de documents, un ensemble de requêtes (représentations du besoin en information), un ensemble de jugements de pertinence (qui indique pour chaque requête, les documents qui satisfont ce besoin en information et ceux qui ne le satisfont pas). Les évaluations sont typiquement lancées par un processus *batch* ; le système à évaluer retourne un nombre pré-spécifié de documents en réponse à chaque requête, sans aucune interaction avec l'utilisateur.

En outre, dans le modèle de Cranfield le protocole d'évaluation est orienté topics (requêtes) et non utilisateur. L'efficacité de l'exécution est mesurée en utilisant un ensemble de métriques *thématiques* dérivées du nombre de réponses (c'est à dire, en termes de documents retournés) pertinentes qui ont été trouvés. Les tests effectués ne prennent pas en considération ni le contexte dans lequel se fait la recherche, ni la perception de la pertinence de l'utilisateur dans ce contexte, ni de la diversité des centres d'intérêts de l'utilisateur. Il est donc difficile de déterminer des collections de tests traduisant l'aspect subjectif de la notion de pertinence et des centres d'intérêts des utilisateurs. En effet, le modèle de Cranfield ne traite pas les besoins dynamiques en information mais les besoins sont considérés comme des concepts statiques entièrement reflétés par la requête. La conclusion est que le processus d'évaluation en mode batch du modèle de Cranfield n'est pas approprié à l'évaluation des systèmes interactifs d'accès personnalisé.

4.3.3 Les protocoles d'évaluation pour l'accès personnalisé

Des différents éléments abordés dans la section précédente, force est de constater qu'il n'y a actuellement aucune tâche de personnalisation dans TREC. Aucune collection de test standard n'a été construite à notre connaissance pour évaluer l'efficacité de l'accès personnalisé à l'information. De telles collections contiendraient divers éléments du contexte liés à l'utilisateur directement (historique de la recherche, centres d'intérêt, expertise etc.) ou à la session de recherche (but de la recherche, tâche, etc.).

En plus de l'absence de collections de tests, la recherche dans ce domaine est confrontée à l'inexistence de méthodologies formelles, de mesures standards d'évaluation de l'adéquation des profils appris aux centres d'intérêts de l'utilisateur, ni l'existence de système référentiel. Il est d'autant plus difficile de réaliser des scénarios d'évaluations objectifs en intégrant la dimension de l'utilisateur dans le processus d'accès. Principalement, en raison du caractère subjectif des utilisateurs, des contextes différents de recherche, de la perception de la pertinence relative des utilisateurs dans un même contexte, la définition d'une méthodologie formelle d'évaluation reste une problématique majeure dans le domaine de la recherche d'information personnalisée.

Nous abordons dans ce qui suit les éléments nécessaires à la mise en place de ce type d'évaluation : les principales mesures d'évaluation ayant émergé dans les travaux de référence sur l'évaluation de systèmes d'accès interactif à l'information ; les approches pour l'élaboration

de collection de test et les scénarios d'évaluation envisageables.

4.3.3.1 Les mesures d'évaluation

Différentes mesures d'évaluation ayant été proposées dans le cadre des travaux sur la recherche des systèmes interactifs. Ces mesures peuvent être également employées pour l'évaluation d'un système d'accès personnalisé à l'information [174].

1. *la mesure RR (Relative Relevance).*
   La mesure RR [19] a pour objectif de considérer différents types de pertinence (pertinence non binaire) dans l'évaluation de l'efficacité d'un système d'accès contextuel à l'information. Cette mesure quantifie le degré de concordance entre les types de jugement de pertinence émis dans le cas de deux ensembles de jugements (soit R_1 et R_2) associés à une même liste de documents qui constitue les résultats d'une session de recherche. En pratique, R_1 correspond généralement aux scores de pertinence algorithmique retournés par un SRI et R_2 à des scores de pertinence contextuelle correspondant à un type de pertinence donné : situationnelle si elle est exprimée par un utilisateur, thématique si elle est exprimée par un assesseur etc. La valeur de corrélation entre R_1 et R_2 est généralement calculée en utilisant une mesure du cosinus ; elle quantifie globalement, la capacité du système à prédire le type de pertinence contextuelle considéré.
   A la différence de la mesure classique de précision, cette mesure permet de considérer les différents types de pertinence ; néanmoins, elle pose un problème lors de l'évaluation comparative entre différents algorithmes de recherche voire entre différents SRI [18]. En effet les scores de pertinence algorithmique ne sont pas étalonnés à la même échelle entre différents SRI, ce qui rend la comparaison de mesures RR non significative.
2. *les mesure CG (Cumulative Gain) et DCG (Discount Cumulative Gain)*
   Les mesures CG et DCG [82; 83] sont des mesures orientées position définies dans le contexte d'une pertinence graduelle et dont l'objectif est d'estimer le gain de l'utilisateur en terme de pertinence cumulée en observant les documents situés jusqu'à un rang donné. Ces mesures sont définies comme suit :

$$CG[i] = \begin{cases} G[1], si\ i = 1 \\ CG[i-1] + G[i],\ sinon \end{cases} \tag{4.8}$$

   où $G[i]$ est la valeur de pertinence associée au document de rang i.

$$CG[i] = \begin{cases} G[1], si\ i = 1 \\ CG[i-1] + G[i]/log_q,\ sinon \end{cases} \tag{4.9}$$

   Comparativement à la mesure CG, la mesure DCG permet d'atténuer le gain de pertinence apporté par un document en fonction du rang associé. Ceci rejoint en effet l'hypothèse évidente que plus le rang d'un document est élevé, moins il est probable que l'utilisateur l'examine et donc moins il est à l'origine d'un gain effectif de pertinence.

3. *La mesure GRP (Generalised Recall and Precision)*
   La mesure GRP [82] est également une mesure orientée position qui généralise les mesures classiques de rappel et précision en considérant une pertinence graduelle. Le rappel généralisé (GR) et la précision généralisée (GP) sont calculés comme suit :

$$GP = \sum_{d \in R} r(d) / |R| \tag{4.10}$$

$$GR = \sum_{d \in R} r(d) / \sum_{d \in D} r(d) \tag{4.11}$$

   où R est l'ensemble des documents retournés par le SRI, D est l'ensemble des documents de la collection, $r(d)$ est la valeur de pertinence graduelle associée au document d.
   De manière analogue aux mesures classiques de rappel/précision, ces mesures offrent la possibilité d'être agrégées pour plusieurs requêtes ou plusieurs niveaux de rappel et donnent ainsi la possibilité de tracer des courbes de performances.

4.3.3.2 Collection de test

La littérature (relativement récente) fait état de deux principales démarches de construction de collections de test dans le cadre de l'accès personnalisé à l'information :

1. réutilisation des collections de test de TREC (documents, requêtes et jugements de pertinence) puis leur augmentation par des éléments du contexte. Ces éléments, tel que l'historique des interactions, sont extraits à partir des interactions d'utilisateurs effectifs interrogeant la base TREC à l'aide de requêtes TREC [159]. Le référentiel d'évaluation étant disponible, les mesures agrégées de rappel/précision sont alors exploitées pour évaluer les différences de performances entre le scénario de recherche basique (ne tenant pas compte du contexte) et le scénario de recherche contextuelle.
2. construction de collections de test en menant une campagne d'évaluation avec des utilisateurs réels : c'est le protocole adopté par la plupart des travaux. Dans ce cas, la majorité des collections utilisées sont extraites à partir du *web* [134; 65; 113; 165; 180].

Un ensemble d'utilisateurs est identifié (étudiants, clients, etc.) et un ensemble de requêtes est construit. La démarche utilisée, de manière générale, est la suivante :

* x utilisateurs soumettent n requêtes au SRI.
* Chaque utilisateur juge les k premiers pour chaque requête.
* Collecter un volume de données test issu du croisement des interactions avec le SRI lié à l'évaluation des résultats de la requête (jugements, lecture, sauvegarde, etc.) pour chaque utilisateur spécifique et pour chaque requête. L'ensemble des documents jugés constitue la collection de référence.

4.3.3.3 Scénarios d'évaluation d'un SRIP

Divers travaux ont tenté de mettre en place un cadre d'évaluation approprié aux SRIs personnalisés. Il en ressort que l'objectif d'un tel protocole d'évaluation est de mesurer l'efficacité de la méthode d'apprentissage (construction et évolution) du profil utilisateur, et évaluer l'impact de l'intégration du profil utilisateur dans le processus d'accès sur les performances de recherche. De ce fait, tout protocole d'évaluation doit répondre à deux exigences :

1. **Valider l'approche de personnalisation** en mesurant l'adéquation du profil utilisateur ainsi que l'efficacité de la méthode de construction du profil utilisateur.
2. **Tester les paramètres de l'approche de personnalisation** à travers la comparaison des performances du SRIP obtenus avec l'intégration du profil de l'utilisateur et ceux obtenus sans son intégration.

Ainsi, de manière générale les scénarios d'évaluation s'effectuent selon la démarche suivante :

– **Etape 1** : Evaluer la qualité des profils appris.
  Lors de cette étape, la qualité du profil se traduit par son adéquation avec les centres d'intérêts effectifs de l'utilisateur. Pour cela, un découpage de la collection de test est effectué en deux sous-collections : une sous-collection pour l'apprentissage du profil utilisateur et une sous-collection pour les tests à effectuer. Et ensuite, ces tests peuvent être effectués en utilisant des mesures quantitatives.
  Ces mesures permettent de quantifier le degré de précision des profils construits relativement aux annotations explicites des utilisateurs [35; 57]. En outre, cette étape peut inclure des tests pour évaluer l'efficacité de l'algorithme d'apprentissage du profil. Dans ce cas, des mesures comparatives entre plusieurs algorithmes [134] peuvent être utilisées où des mesures de convergence de l'algorithme [113].
– **Etape 2** : Validation de l'accès personnalisé.
  L'objectif de cette étape est de tester l'amélioration des performances de la recherche. Les scénarios expérimentaux consistent, de manière classique, à comparer les performances de recherche d'un moteur de recherche classique (sans intégration du profil) et du moteur de recherche personnalisé proposé intégrant le profil de l'utilisateur [113; 165; 65]. Dans le cas de l'utilisation de mesures agrégées de rappel/précision, un référentiel est généralement construit sur la base de l'ensemble des documents pertinents jugés par l'ensemble

des utilisateurs pour chaque requête. L'utilisation de mesures orientées rang évite l'utilisation d'un tel référentiel.

Nous présentons dans ce qui suit, un exemple de deux travaux de référence dans le domaine ayant suivis la méthodologie d'évaluation que nous venons de décrire.

1. Dans les travaux de [113] la personnalisation consiste en la désambiguïsation de la requête de l'utilisateur en se basant sur le profil de l'utilisateur. Pour valider leurs approches, ils effectuent les expérimentations se déroulant avec 7 utilisateurs. Chaque utilisateur soumet *n* requêtes à un *Google web Directory* en identifiant les catégories associées pour chacune des ces requêtes. Pour établir la collection de test, chaque requête est exécutée selon 3 modes :

   (a) *Mode de base* : sans aucune spécification des catégories par l'utilisateur ;

   (b) *Mode semi automatique* : avec spécification par l'utilisateur des catégories identifiées par le système ;

   (c) *Mode automatique* : avec spécification des catégories automatiquement par le système.

   La collection de test référentiel est ensuite obtenue en regroupant pour chaque paire (utilisateur, requête) l'union de l'ensemble des documents jugés par l'utilisateur pour les 3 modes de soumission de la requête. L'évaluation des profils appris passe par le test de l'efficacité des algorithmes utilisés pour la construction de ce profil en comparant les différents résultats obtenus par chacun des algorithmes.

   Pour tester les performances de l'algorithme de construction du profil, ils augmentent à chaque apprentissage la taille des données (i.e., la taille de l'historique de recherche) en appliquant la stratégie de la *k*-fold Cross Validation, où *k* est positionné à 10. Ils découpent ainsi la collection de test en 10 sous collections : 9 sous collections pour l'apprentissage et la 10^{ieme} pour le test. Puis, ils répètent l'expérimentation 10 fois. A chaque *ième* expérimentation, *ième* sous collection est utilisée pour le test.

2. Les secondes expérimentations que l'on cite sont ceux des travaux de [165]. Le scénario général d'évaluation s'effectue avec 6 utilisateurs (étudiants de l'université du Kansas) durant 6 mois. Chaque utilisateur soumet 45 requêtes à Google. Durant cette période, ils collectent pour chaque paire (utilisateur, requête) les 10 premiers résultats sélectionnés par l'utilisateur. Puis, ils découpent cette collection en deux sous ensembles :

   * Du résultat des 40 requêtes, ils forment la collection pour l'apprentissage du profil.
   * Du résultat des 5 requêtes restantes, ils forment la collection de test.

   Pour évaluer les profils appris, ils effectuent une série de tests en faisant varier la taille des collections d'apprentissage pour la construction du profil. Les profils utilisateurs sont créés sur la base de 5, 10, 20, 30, puis 40 requêtes. Ils construisent également des profils avec 30 requêtes et un nombre de 20 concepts issus d'ODP.

Par la suite, lors de l'évaluation de l'impact de l'intégration du profil dans le processus d'accès, la collection référentielle est obtenue par l'union des documents jugés pertinents par l'utilisateur pour l'ensemble des requêtes. Ils mesurent l'exactitude des profils construits en effectuant des comparaisons statistiques entre le rang calculé par le système (sur la base de similarité avec le profil) et celui obtenu par Google pour chacun des 10 premiers résultats sélectionnés par l'utilisateur.

4.4 Conclusion

Nous avons présenté dans la première partie de ce chapitre les principaux modèles d'accès personnalisé à l'information. Notre synthèse fait ressortir d'emblée que les approches et techniques associées puisent largement des acquis reconnus dans les domaines de la RI. Nous avons examiné un échantillon représentatif de ce type de systèmes qui ont été développés jusqu'ici. Nous les avons regroupés selon des modèles mettant en évidence leurs principales approches d'accès personnalisé. Ils constituent un ensemble d'approches possibles de personnalisation, qui cependant ne soutient pas une comparaison directe entre eux car ils tendent généralement à réaliser différents objectifs de la personnalisation.

En effet, l'adaptation cycle de vie de la requête dans le processus d'accès s'effectuée principalement à l'un des différents niveaux : sélection des sources d'information, reformulation de requête, sélection de l'information et réordonnancement des résultats.

– La reformulation de requête a pour objectif d'introduire dans la structure de la requête les termes issus du profil de l'utilisateur ; c'est la technique la plus largement répandue.
– La sélection adaptée de l'information, promue par de récents travaux, évoque la contextualisation de la fonction de pertinence en définissant des paramètres issus du profil de l'utilisateur.
– L'adaptation consiste généralement à réordonnancer les résultats en tenant compte de critères descriptifs de l'utilisateur.

La seconde partie de ce chapitre est consacrée aux approches d'évaluation des systèmes d'accès personnalisé à l'information. Des différents éléments abordés dans cette section, il en ressort deux points importants :

– les campagnes d'évaluation standard largement utilisées RI tel que TREC, ne sont pas adaptées à la RI personnalisée. Ces protocoles d'évaluation sont centrés requête et non utilisateur.
– l'évaluation orientée « utilisateur » est une composante primordiale dans le cadre de l'accès personnalisé à l'information. Les objectifs d'une telle évaluation sont de mesurer l'adéquation des profils utilisateur construits par le système avec les centres d'intérêts effectifs de l'utilisateur ; ainsi que l'impact de l'intégration de ce profil, dans le processus d'accès, sur les performances de recherche.

Il est d'autant plus difficile de réaliser des scénarios d'évaluations objectifs en intégrant la

dimension de l'utilisateur dans le processus d'accès. Principalement, en raison du caractère subjectif des utilisateurs, des contextes différents de recherche, de la perception de la pertinence relative des utilisateurs dans un même contexte, la définition d'une méthodologie formelle d'évaluation reste une problématique majeure dans le domaine de la recherche d'information personnalisée.

Deuxième partie

Spécification et évaluation d'un modèle de RI personnalisé

Motivations

L'état de l'art présenté dans la première partie de cette thèse a permis de cerner le domaine de l'accès personnalisé à l'information dans sa globalité. Il en ressort différents aspects dont les plus importants sont les suivants :

1. L'efficacité du processus d'accès à l'information dépend fortement de la précision des profils utilisateurs modélisés : les travaux s'orientent vers la construction de profils selon des algorithmes et des techniques basées sur l'évaluation implicite de la pertinence. Les sources d'information utilisées sont extraites à partir des données issues des interactions de l'utilisateur avec le SRI et organisées selon des structures ensemblistes basées sur les vecteurs de termes ou sémantiques issues d'une ontologie de concepts. L'évolution est généralement abordée comme une procédure inhérente à la construction à travers la mise à jour des structures des vecteurs représentant le profil ou l'intégration de nouvelles catégories sémantiques.
2. Fournir un accès personnalisé ne peut s'effectuer sans inclure le modèle utilisateur, en l'occurrence son profil, comme composante à part entière dans l'une ou plusieurs phases du cycle de vie d'une requête : reformulation, calcul d'un score personnel, réordonnancement des résultats. On note que, dans la majorité des travaux de la littérature, les données du profil sont exploitées principalement pour réordonner les résultats de recherche selon des techniques basées sur la combinaison et/ou la fusion des scores de pertinence.
3. Absence de collections de test standard pour une tâche d'accès personnalisé à l'information. La plupart des travaux du domaine conduisent des expérimentations dans le cadre de campagnes d'évaluation locales menées avec des utilisateurs effectifs, coûteuse en temps. Récemment, des travaux précurseurs s'orientent vers la réutilisation de collections TREC en intégrant les facteurs issus de l'utilisateur.

Inscrits dans la lignée de ces travaux, les travaux présentés dans cette thèse investissent les trois questions critiques posées par un processus d'accès personnalisé à l'information : **modélisation de l'utilisateur**, **modélisation de l'accès** et **évaluation du modèle d'accès**.

Dans ce qui suit, nous présentons les grandes lignes de chacun des aspects de notre contribution en mettant en exergue sa spécificité relativement aux travaux du domaine.

1. Au niveau de la **modélisation du profil de l'utilisateur**. Notre intuition repose sur le fait que le corpus ou collection de documents, géré par le SRI, représente un espace informationnel, dont la perception par les utilisateurs est définie à travers leurs interactions. De ce fait, l'exploitation de ces interactions permet de renforcer ou d'affaiblir des liens d'association sémantiques entre information au regard de ces interactions. Les synergies qui sont ainsi créées vont nous amener à densifier un espace de représentation de l'utilisateur en favorisant l'émergence de groupes de concepts correspondant à ses centres d'intérêts.

Comparativement aux travaux du domaine, notre objectif est :

(a) D'inférer (et donc de distinguer) les données caractéristiques des centres d'intérêts de l'utilisateur à partir des données de comportement issues des sources d'information. En outre, plutôt que d'exploiter l'importance intrinsèque des termes dans les documents, on propose de décliner l'importance relative des termes selon le profil de l'utilisateur.

(b) De faire évoluer les centres d'intérêts de l'utilisateur en scrutant la variation dans les sujets des requêtes d'une session à une autre.

A cet effet, notre approche de modélisation est caractérisée par :
– La définition d'un profil utilisateur à deux dimensions : « *Historique de recherche* » et « *Centres d'intérêts* ».
– Le procédé de construction du profil utilisateur applique des opérateurs d'agrégation afin de décliner le degré d'importance des informations collectées des interactions de l'utilisateur comparativement à l'ensemble de l'historique [176].
– Le procédé d'évolution du profil utilisateur repose sur l'interaction entre ses centres d'intérêts et l'historique de ses recherches, sans utilisation d'autres ressources tels que les ontologies ou classifieurs de concepts. Une méthode statistique est déployée pour cela afin de détecter, en cours du temps, les différents changements dans les centres d'intérêts [172].

2. Au niveau du **modèle d'accès personnalisé à l'information**. Nous appréhendons l'évaluation de la pertinence comme un problème décisionnel lié à la présentation ou non d'un document en fonction de plusieurs critères :

    * La pertinence du document relativement à la requête,
    * la pertinence du document relativement aux centres d'intérêts de l'utilisateur
    * et l'adéquation des centres d'intérêts de l'utilisateur vis-à-vis de la requête.

    Ainsi, les centres d'intérêts de l'utilisateur sont exploités, lors de l'évaluation de la requête, comme une composante à part entière dans le modèle d'accès personnalisé et non comme une source de définition d'heuristiques ou techniques permettant la réécriture de la requête et/ou la fonction d'appariement [173]. A notre connaissance, aucune étude n'a tenté d'appréhender le problème de l'accès personnalisé comme un processus décisionnel, intégrant le profil utilisateur comme composante du modèle de recherche.

    Plus précisément, nous formalisons ce problème à l'aide d'un modèle inférentiel, basé sur les diagrammes d'influence (DI), extension des réseaux Bayésiens (RB). Le modèle permet de formaliser le raisonnement lié à la prise de décision quant à la pertinence des documents compte tenu du profil de l'utilisateur d'une part et de la requête d'autre part [201].

3. Au niveau de **l'évaluation des performances** d'un processus d'accès personnalisé à l'information. Nous proposons un cadre d'évaluation adapté à l'accès personnalisé à l'information en réutilisant les ressources TREC [202; 173]. Comparativement aux travaux ayant suivi cette voie, notre protocole d'évaluation s'en distingue par :
   – La dérivation des facteurs liés à l'utilisateur, notamment ses centres d'intérêts, à partir de la collection de test et non d'utilisateurs effectifs.
   – L'utilisation d'un procédé de validation croisée pour la construction des centres d'intérêts à partir de documents pertinents manuellement annotés de domaines d'intérêts généraux.

Nous introduisons de manière sommaire dans ce qui suit les grandes lignes de notre approche.

Notre démarche

La figure 4.7 présente les principaux éléments de notre démarche de personnalisation. Les composantes 1 et 2 concernent le processus de modélisation du profil. La composante 3 est le processus décisionnel d'accès personnalisé à l'information.

1. **L'acquisition des données utilisateur**

   Ces données correspondent aux documents jugés pertinents explicitement et/ou implicitement par l'utilisateur. L'acquisition *implicite* est réalisée en observant le comportement de l'utilisateur face aux résultats de recherche (*Lecture, Sauvegarde, Impression*). A cet effet, nous avons mis en place un outil nommé $\mathcal{W}eb\ Cap$[7] capable d'inférer de ces comportements l'intérêt de l'utilisateur pour les résultats de recherche traduisant sa perception de la pertinence.

2. **Processus de modélisation du profil utilisateur**

   Correspond à l'inférence des différents centres d'intérêts de l'utilisateur à partir des informations collectées dans son historique de recherche. L'émergence de ces centres est obtenue en détectant les différents changements d'intérêts de l'utilisateur lors de ses sessions de recherche successives. Notre stratégie se déroule en deux étapes :

   (a) *Construction* et *évolution* de l'historique de recherche à partir des données collectées lors de ses différentes sessions de recherche,

   (b) *dérivation* des paquets d'information représentant des *contextes* d'intérêts à court terme à partir de l'historique de recherche, puis *évolution* des centres d'intérêts sur la base de corrélation entre ces contextes pour mettre à jour la librairie de centres d'intérêts de l'utilisateur.

3. **Modèle d'accès personnalisé décisionnel**

   La définition d'un modèle décisionnel pour l'accès personnalisé à l'information basé sur les Diagrammes d'Influence. Dans ce modèle, le calcul du score pertinence d'un document relativement à une requête émise par l'utilisateur, est fondé sur une fonction cumulative permettant d'estimer l'utilité de présenter à l'utilisateur des documents et ce, compte tenu de sa requête et de ses centres d'intérêts spécifiques.

Les chapitres de cette partie abordent en détail nos propositions.

[7] pour *web* capture, analogiquement à la capture de l'activité sur le *web*

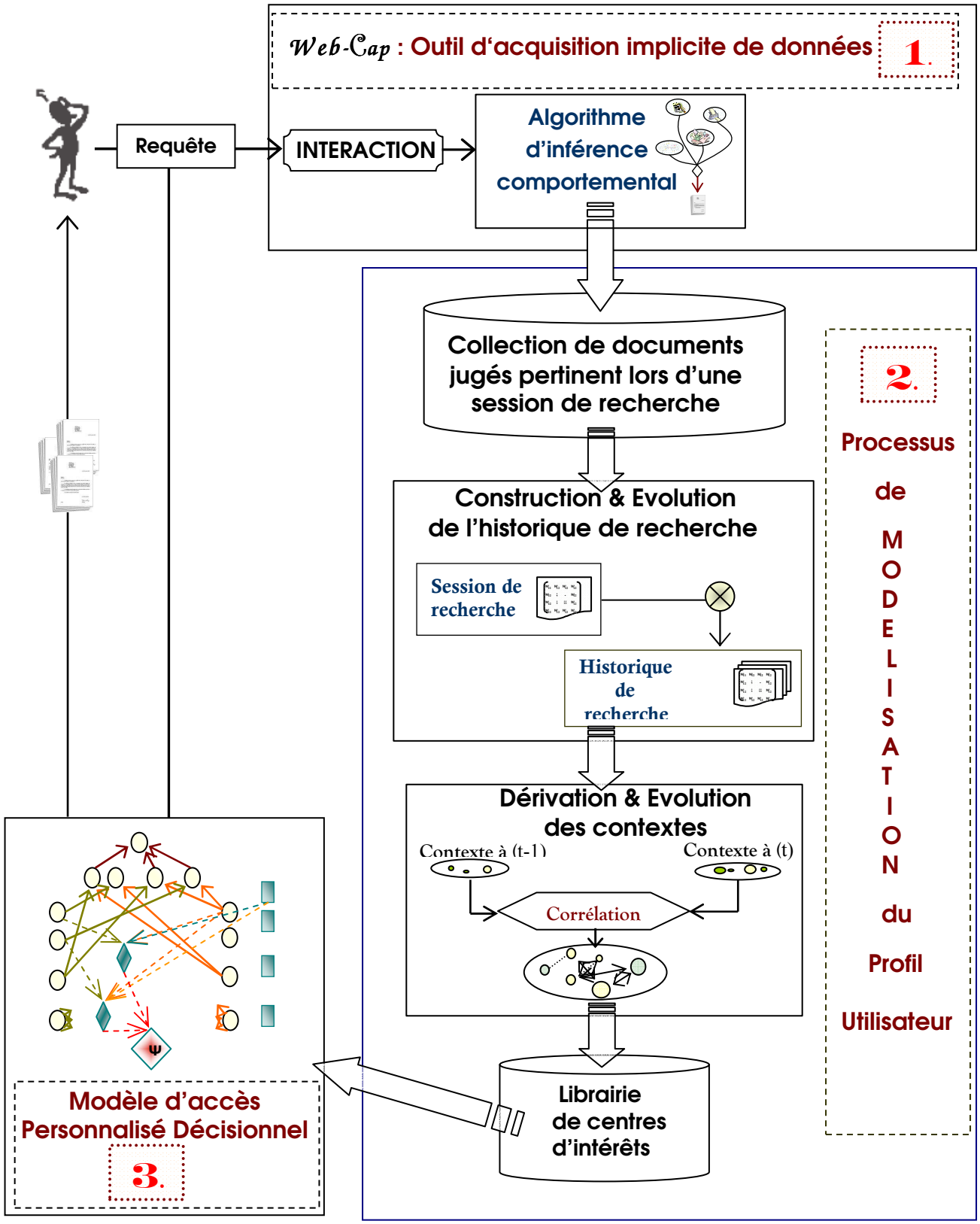


Figure 4.7 — Architecture générale de notre contribution

5 Profil Utilisateur : Interaction, Inférence et Evolution

5.1 Introduction

Ce chapitre est consacré à la modélisation des dimensions informationnelles descriptives du profil utilisateur. Dans notre approche, on gère une librairie de centres d'intérêts inférés automatiquement à partir de l'historique de recherche de l'utilisateur. Cet historique est vu comme une source d'information, évoluant lors des différentes sessions de recherche de l'utilisateur, à partir desquelles on fait émerger les centres d'intérêts de l'utilisateur. Plus précisément, notre procédé de construction du profil se décline en un cycle comportant deux principales étapes :

– La première étape consiste à construire puis faire évoluer l'historique de recherche de l'utilisateur avec le SRI par agrégation des informations collectées à partir de ses sessions de recherche successives.
– La seconde étape a pour but de construire puis faire évoluer les centres d'intérêt de l'utilisateur en se basant sur la dimension « *Historique de recherche*». Plus précisément, on détermine des périodes d'apprentissage qui définissent des jalons pour l'extraction de centres d'intérêt à court terme, qualifiés de *contextes d'usage*, à partir des informations agrégées dans l'historique de recherche. L'évolution de la dimension « *Centres d'intérêts* » est alors basée sur une mesure de corrélation des rangs qui évalue le degré de changement entre contextes d'usage associés à des périodes successives.

Nous donnons dans la section 5.2 les définitions de base des éléments constituant le profil utilisateur ainsi que les notations associées utilisées dans notre approche. Dans la section 5.3, nous introduisons un exemple de collection pour illustrer notre approche de modélisation du profil utilisateur. Dans la section 5.4, nous abordons notre approche d'agrégation pour la construction de la dimension « *Historique de recherche* ». Dans la section 5.5, on présente notre approche statistique pour l'inférence de la dimension « *Centres d'intérêts* » à partir de l'historique de recherche. Dans la section 5.6 nous décrivons une approche d'acquisition implicite de documents pour la construction de la dimension « *Historique de recherche* ». Nous conclurons ce chapitre par une synthèse des avantages de notre approche.

5.2 Définitions et Notations

Nous introduisons dans cette section les définitions des notions suivantes : *profil utilisateur*, *session de recherche*, *historique de recherche*, *contexte d'usage* et *centre d'intérêt*.

Soit $\phi = \{D, T\}$ l'univers des informations de base manipulées par le modèle, tels que :

- $D = \cup_{j=1}^{n} d_j$ est l'ensemble des documents de la collection et n le nombre total de documents.
- $T = \cup_{i=1}^{m} t_i$ est l'ensemble des termes indexant les documents de D et m le nombre total de termes d'indexation.

1. Un **profil** est une structure informationnelle représentant l'utilisateur à travers deux dimensions corrélées, évoluant dans le temps, noté $U = (H^s, C^s)$. tels que H^s représente l'historique de recherche de l'utilisateur jusqu'à l'instant s et C^s représente la librairie de ses centres d'intérêt inférés jusqu'à l'instant s.
2. Une **session de recherche** est décrite par une requête et un ensemble de documents associés, jugés explicitement ou implicitement pertinents par l'utilisateur.

   Soit q^s une requête soumise par un utilisateur u à la session de recherche se déroulant à l'instant s, notée S^s, et D^s l'ensemble des documents pertinents pour l'utilisateur durant cette session. On note alors $R_u^s = \cup_{s_0 \ldots s_{(s-1)}} D^s$ l'ensemble des documents déjà *visités* et jugés pertinents par l'utilisateur lors des sessions de recherche passées depuis l'instant s_0. On note $T(R_u^s)$ le sous ensemble de termes indexant les documents de R_u^s.
3. Un **historique de recherche** est l'ensemble des informations collectées de l'utilisateur au cours de ses sessions de recherche jusqu'à l'instant s.
4. Un **centre d'intérêt** est l'ensemble des besoins en information récurrents de l'utilisateur regroupés dans la librairie C^s. Chaque centre d'intérêt est représenté par un vecteur de termes pondérés.
5. Alors qu'un **contexte d'usage** traduit un besoin en information à court terme exprimé sur une courte période d'interactions avec le SRI. Chaque contexte d'usage est également représenté par un vecteur de termes pondérés.

5.3 Exemple illustratif

La collection de documents que nous présentons ci-dessous servira à illustrer notre approche de modélisation du profil utilisateur. Pour chaque phase du processus de modélisation, nous étayons nos propos en présentant le résultat de notre approche sur cet exemple.

Supposons que l'on dispose de la collection de documents suivante :

$D = \{D_1,\ D_2,\ D_3,\ D_4,\ D_5,\ D_6,\ D_7,\ D_8\}$ tel que $T = \{t_1,\ t_2,\ t_3,\ t_4,\ t_5,\ t_6,\ t_7\}$ est l'ensemble des termes indexant ces documents. La fréquence d'apparition de chaque terme $t_i \in T$ dans chaque document $d_j \in D$ est donnée comme suit :

$d_1 = \{42t_1,\ 11t_3,\ 6t_5,\ 8t_7\}, d_2 = \{3t_1,\ 15t_2,\ 20t_3,\ 17t_5,\ 3t_6\}, d_3 = \{5t_3,\ 10t_4,\ 24t_6\}$, $d_4 = \{2t_1,\ 4t_2,\ 13t_4,\ 4t_5,\ 26t_7\}, d_5 = \{7t_3,\ 33t_6,\ 18t_7\}, d_6 = \{15t_1,\ 3t_2,\ 2t_4,\ 11t_5,\ 3t_7\}$, $d_7 = \{6t_1,\ 7t_2,\ 23t_3,\ 1t_4,\ 45t_5,\ 4t_6,\ 8t_7\}, d_8 = \{9t_2,\ 2t_3,\ 14t_5,\ 6t_6,\ 2t_7\}$.

Dans cet exemple, on suppose que les documents ayant été jugés pertinents par l'utilisateur u durant s sessions de recherche sont formés par le sous-ensemble de document $R_u^s = \{D_1, D_2, D_3, D_4, D_5, D_6\}$.

Le tableau 5.1 représente un récapitulatif des poids associés $w_{i,j}$ de chaque termes $t_i \in T$ indexant les documents $D_j \in R_u^s$, obtenu par application de la formule du $Tf \star Idf$ normalisée.

	t_1	t_2	t_3	t_4	t_5	t_6	t_7
w_{i1}	0.23	0	0.51	0	0.14	0	0.21
w_{i2}	0.13	0.68	0.65	0	0.16	0.17	0
w_{i3}	0	0	0.42	0.66	0	0.27	0
w_{i4}	0.12	0.43	0	0.75	0.13	0	0.33
w_{i5}	0	0	0.45	0	0	0.31	0.28
w_{i6}	0.16	0.41	0	0.43	0.15	0	0.18

Tableau 5.1 — Poids des termes de $T(R_u^s)$

5.4 Construction de l'historique de recherche

Cette section introduit le principe de construction de l'historique de recherche de l'utilisateur sur la base des informations collectées à partir de ses différentes sessions de recherche.

L'idée de base de notre approche est d'exploiter le comportement passé de l'utilisateur pour déduire son comportement futur. En effet, l'activité de recherche de l'utilisateur traduit son intérêt pour un ou plusieurs centres d'intérêts. Notre objectif est de traduire cet intérêt en capitalisant l'importance des informations ayant été jugées pertinentes par l'utilisateur. Ces informations sont contenues dans les différentes sessions de recherche, que l'on agrège pour former l'historique de toutes les recherches de l'utilisateur. Dans le but d'exploiter ces informations, on propose dans ce qui suit l'utilisation de matrices pour la représentation d'une session de recherche et de l'historique des interactions.

5.4.1 Représentation d'une session de recherche

Chaque session de recherche utilisateur est exprimée par un tuple d'information identifiée comme suit :

session := < texte de requête > [document sélectionné]*

Afin d'exploiter ces informations, nous les représentons sous forme matricielle : *Document-Terme* notée $D^s x T^s$ où T^s est l'ensemble des termes qui indexent les documents D^s (T^s est une partie de l'ensemble des termes représentatifs des documents préalablement jugés pertinents noté $T(R_u^s)$). Chaque ligne de la matrice S^s représente un document $d \in D^s$, chaque colonne représente un terme $t \in T^s$.

Dans le but d'affiner la représentation Document-Terme, on propose de décliner l'importance d'un terme relativement au profil de l'utilisateur dans le schéma de pondération terme-document. A cet effet, on calcule pour chaque terme t dans un document d à l'instant s, un coefficient d'importance $CPT^s(t,d)$ qui traduit la pertinence relative d'un terme compte tenu des jugements de pertinence qu'il a émis et qui sont supposés être des indicateurs de son centre d'intérêt courant [172].

L'expression de ce coefficient est fondée sur l'hypothèse qu'un terme est d'autant plus important pour l'utilisateur qu'il cooccurre avec les termes qui lui sont *familiers*, en ce sens qu'ils sont présents dans des documents déjà jugés. Les dépendances entre termes associés à des documents préalablement jugés sont vues comme des règles d'association [112].

Définition 1. *Le coefficient d'importance d'un terme t dans un document d à l'instant s noté* $CPT^s(t,d)$ *est défini comme suit* :

$$CPT^s(t,d) = \frac{w_{(t,d)}}{dl} * \sum_{t' \neq t, t' \in T(R_u^s)} cooc(t,t') \tag{5.1}$$

$w_{(t,d)}$ est le poids du terme t dans le document d calculé selon le schéma classique $Tf \star Idf$, dl est la longueur du document d, $cooc(t,t')$ est le degré de confiance de la règle $(t \rightarrow t')$ quantifié à l'aide de la mesure EMIM (Expected Mutual Information Measure), $cooc(t,t') = P(t,t') log \frac{P(t,t')}{P(t)P(t')}$, $P(t,t')$ est la proportion de documents contenus dans R_u^s indexés à la fois par les termes t et t', $P(t)$ est la proportion de documents contenus dans R_u^s indexés par le terme t.

$S^s(d,t)$ est alors ainsi construit :

$$S^s = (CPT^s)^{Transp} \tag{5.2}$$

Où *Transp* est l'opérateur transposée de matrice.

5.4.2 Agrégation des sessions de recherche

L'historique des interactions de l'utilisateur est représenté par une matrice notée H^s de dimensions $R_u^s * T(R_u^s)$. Cette matrice est construite de manière incrémentale, en ce sens qu'elle est mise à jour à chaque session de recherche en y reportant, par agrégation, les informations issues de la matrice S^s. A cet effet, on propose de définir un opérateur d'agrégation qui combine pour chaque terme son poids classique dans le document et ses poids atténués par les coefficients de pertinence calculés lors des sessions de recherche passées.

Définition 2. *L'opérateur d'agrégation des sessions de recherche, noté $\oplus$, est défini comme suit :*

$H^0(d,t) = S^0(d,t)$

$$H^s(d,t) = H^{s-1}(d,t) \oplus S^s(d,t) = \begin{cases} \alpha * w_{(t,d)} + \beta * S^s(d,t) \; si \; d \notin H^{(s-1)} et \; d \in S^{(s)} \\ \alpha * H^{s-1}(d,t) + \beta * S^s(d,t) \; si \; d \in H^{(s-1)} et \; d \in S^{(s)} \\ H^{s-1}(d,t) \; sinon \end{cases} \quad (5.3)$$

$(\alpha + \beta = 1)$, $s > s_0$

Où, $w_{(t,d)}$ est le poids du terme t dans le document d calculé selon le schéma classique $Tf \star Idf$

La définition de l'opérateur $\oplus$ est fondée sur l'hypothèse que les termes associés aux centres d'intérêt de l'utilisateur sont récurrents. L'idée est alors d'affiner les descripteurs des documents déjà jugés par :

- expansion éventuelle avec des termes associés présents dans des documents pertinents,
- combinaison de l'importance classique de ces termes (relativement à la collection de documents) et de leur pertinence relative au profil calculée à l'aide du coefficient $CPT(t,d)$ au cours des sessions de recherche passées.

Les deux paramètres α et β de la fonction d'agrégation (5.3) jouent un rôle important dans la mise à jour de l'historique de recherche. En effet, les poids affectés aux termes indexant ce document varient selon la source d'information que l'on souhaite favoriser. En d'autres termes, si l'on considère que la session de recherche induit plus d'importance dans la représentativité du terme dans le document, il est alors important de quantifier cette information en favorisant le poids de ce terme dans la session par rapport à l'historique. Dans ce cas, on affecte des valeurs à α et β tel que $\alpha < \beta$.

Si l'on suppose que les informations capitalisées dans le temps à travers les interactions de l'utilisateur avec le SRI induites par l'historique $H^{s-1}(d,t)$ sont plus significatives que celles apportées par la session courante, alors $\alpha > \beta$.

Le cas où les valeurs de α et β sont égales, les poids des termes sont atténués relativement à l'historique $H^{s-1}(d,t)$ et la session courante telle que l'importance n'est affectée à aucune

sources d'informations mais tient compte du poids cumulé par les deux sources calculées à l'aide du coefficient $CPT(t,d)$.

5.4.3 Illustration

Dans le but d'illustrer notre approche, on suppose le scénario de recherche suivant effectué par l'utilisateur u à l'instant s_0 sur la collection de documents illustrée précédemment :

* A l'instant s_0 l'utilisateur soumet la requête $q^{s_0} = \{t_2, t_5\}$;
* Soit, $\{D_1, D_4\}$ l'ensemble des documents jugés pertinents par l'utilisateur lors de la session de recherche S^{s_0} ;
* On a alors $T(R_u^{s_0}) = \{t_1, t_2, t_3, t_4, t_5, t_7\}$ l'ensemble des termes indexant les documents d_1 et d_4.

Le tableau 5.2 donne les cooccurrences entre l'ensemble des termes d'indexation $t_i \in T$:

$cooc(t,t')$	t_1	t_2	t_3	t_4	t_5	t_7
t_1	-	3.30	1.75	0.94	6.27	6.27
t_2	3.30	-	0.18	0.94	3.84	1.75
t_3	1.75	0.18	-	-0.41	1.39	1.39
t_4	0.94	0.94	-0.41	-	2.20	2.20
t_5	6.27	3.84	1.39	2.20	-	3.30
t_7	6.27	1.75	1.39	2.20	3.30	-

Tableau 5.2 — Cooccurrences entre les termes

Si l'on considère les cooccurrences représentées par le tableau 5.2, les valeurs calculées du coefficient de pertinence $CPT^s(t,d)$ associé à chaque terme indexant l'ensemble des documents de $R_u^{s_0} = \{d_1, d_4\}$ sont données dans le tableau 5.3 suivant :

$CPT^s(t,d)$	d_1	d_4
t_1	1.09	0.46
t_2	0	0.87
t_3	0.91	0
t_4	0	1.48
t_5	0.57	0.44
t_7	0.80	0.98

Tableau 5.3 — Coefficient de pertinence des termes des documents d_1, d_4

En appliquant le coefficient $CPT^s(t,d)$ donné dans la formule 5.1, on obtient la session de recherche S^{s_0} représentée par la matrice suivante :

$$S^{s_0}(d,t) = \begin{array}{c} \\ d_1 \\ d_4 \end{array} \begin{array}{c} \begin{array}{ccccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 \end{array} \\ \begin{bmatrix} 1.09 & 0 & 0.91 & 0 & 0.57 & 0 & 0.80 \\ 0.46 & 0.87 & 0 & 1.48 & 0.44 & 0 & 0.98 \end{bmatrix} \end{array}$$

Ainsi, selon la fonction d'agrégation 5.3 l'historique de recherche à l'instant s_0 est égal à :

$$H^{s_0}(d,t) = \begin{array}{c} \\ d_1 \\ d_4 \end{array} \begin{array}{c} \begin{array}{ccccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 \end{array} \\ \begin{bmatrix} 1.09 & 0 & 0.91 & 0 & 0.57 & 0 & 0.80 \\ 0.46 & 0.87 & 0 & 1.48 & 0.44 & 0 & 0.98 \end{bmatrix} \end{array}$$

Avec $\alpha = 0.5$ et $\beta = 0.5$.

Supposons à présent, que lors de la session de recherche d'ordre $s_0 + 1$ les documents jugés pertinents sont d_1, d_2 représentés par la matrice suivante :

$$S^{s_0+1}(d,t) = \begin{array}{c} \\ d_1 \\ d_2 \end{array} \begin{array}{c} \begin{array}{ccccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 \end{array} \\ \begin{bmatrix} 1.5 & 0 & 1.08 & 0 & 0.99 & 0 & 1.05 \\ 0.47 & 1.37 & 0.92 & 0 & 0.55 & 0.23 & 0 \end{bmatrix} \end{array}$$

Ainsi, selon la formule 5.3, on obtient l'historique suivant :

$$H^{s_0+1}(d,t) = \begin{array}{c} \\ d_1 \\ d_2 \\ d_4 \end{array} \begin{array}{c} \begin{array}{ccccccc} t_1 & t_2 & t_3 & t_4 & t_5 & t_6 & t_7 \end{array} \\ \begin{bmatrix} 1.29 & 0 & 0.99 & 0 & 0.78 & 0 & 0.93 \\ 0.30 & 1.03 & 0.78 & 0 & 0.36 & 0.20 & 0 \\ 0.46 & 0.87 & 0 & 1.48 & 0.44 & 0 & 0.98 \end{bmatrix} \end{array}$$

Avec $\alpha = 0.5$ et $\beta = 0.5$.

Selon cet exemple on remarque que l'historique $H^{s_0+1}(d,t)$ est mis à jour de la manière suivante :

– Le document d_1 apparaît dans $H^{s_0}(d,t)$ et $S^{s_0+1}(d,t)$. Dans ce cas, nous effectuons une combinaison de l'importance des termes associés présents dans la session courante avec ceux de l'historique.
– Le document d_2 apparaît pour la première fois dans la session courante. Dans ce cas, nous combinons l'importance de ses termes (relativement à la collection de documents) et leur importance relative au profil, calculée à l'aide du coefficient $CPT(t,d)$ au cours de la session de recherche $S^{s_0+1}(d,t)$.

– Le document d_4, n'apparaissant pas dans la session courante, est reporté identiquement de $H^{s_0}(d,t)$ à $H^{s_0+1}(d,t)$.

Si l'on analyse plus finement ces résultats, on remarque que la mise à jour de l'historique induit un changement du degré d'importance du coefficient d'importance des termes de chaque document ajouté.

En outre, cette modification s'effectue selon deux facteurs :

1. *Renforcement du degré d'importance* sur la base des valeurs de cooccurrence entre les termes de la collection, dans le cas où le document appartient à l'historique et a été jugé pertinent dans la session courante. Dans ce cas, plus le terme cooccurre dans la collection plus son coefficient dans l'historique est important.
2. *Atténuation du degré d'importance* sur la base des valeurs des poids associés aux termes indexant le document nouvellement ajouté à l'historique.

5.5 Inférence des centres d'intérêts

Le processus d'inférence des centres d'intérêt est fondé sur une méthode cyclique qui procède en deux étapes :

– La première a pour objet d'extraire à partir de l'historique des interactions un centre d'intérêt candidat qualifié de *contexte d'usage*, qui traduit un besoin à court terme en information.
– L'objectif de la seconde étape est alors d'intégrer le contexte ainsi découvert dans la librairie C^s en respectant l'hypothèse de diversité des centres d'intérêt. Ceci traduit, à juste titre, la phase d'apprentissage des centres d'intérêt qui induit l'évolution du profil.

Ce processus présenté par la figure 5.1 représente des historiques de recherches successives aux instants s et $s+1$. A partir de chaque historique, on extrait un contexte informationnel spécifique, dont on mesure les corrélations pour inférer les centres d'intérêts.

5.5.1 Extraction d'un contexte d'usage

A l'issue d'un cycle d'apprentissage représentant un nombre déterminé de sessions de recherche S^s agrégées dans l'historique H^s, on construit un contexte d'usage courant c^s. Dans notre cas, on considère qu'un cycle d'apprentissage équivaut à la soumission d'une nouvelle requête utilisateur soit, une seule session de recherche.

Un contexte d'usage est représenté par un vecteur de termes pondérés, ordonnés par leur degré de représentativité du contexte ; pour chaque terme $t \in T(R_u^s)$, on calcule le poids associé comme suit :

$$c^s(t) = \sum_{d \in R_u^s} H^s(d,t) \tag{5.4}$$

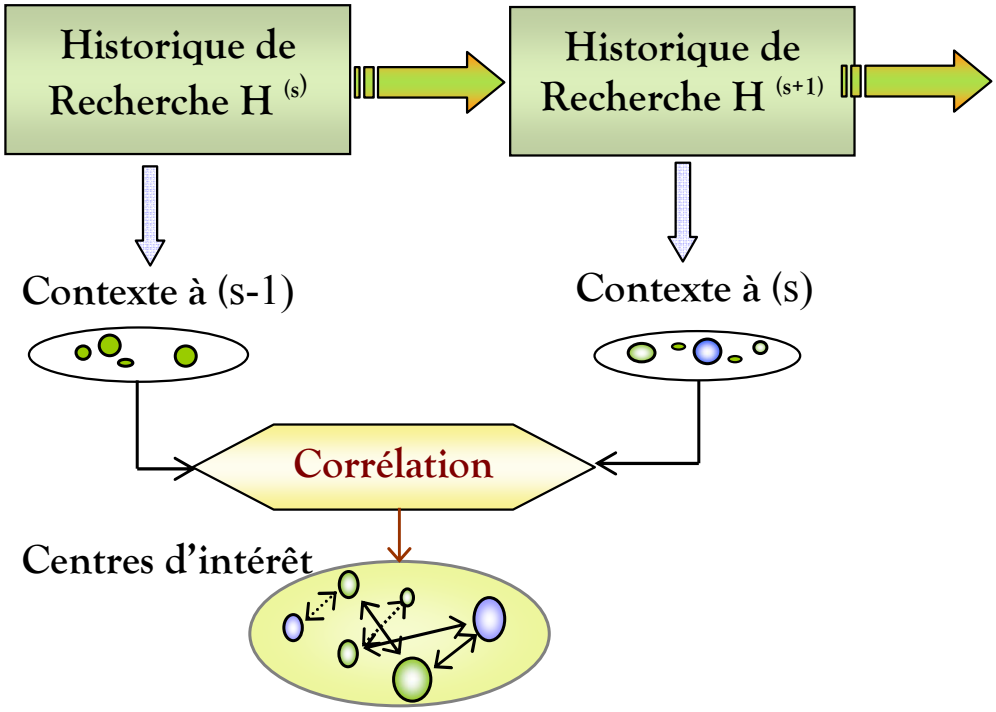


Figure 5.1 — Processus d'inférence des centres d'intérêt

La normalisation des poids de chaque terme du contexte $c^s(t)$ est obtenue par la formule suivante :

$$c^s(t) = \frac{c^s(t)}{\sum_{t \in T(R_u^s)} c^s(t)} \tag{5.5}$$

Un contexte d'usage est ainsi un vecteur extrait à partir de l'historique des interactions en sommant chaque colonne de la matrice associée.

5.5.2 Evolution des centres d'intérêt

La phase d'évolution consiste à exploiter le contexte d'usage extrait précédemment pour évaluer le degré de changement entre les contextes liés aux centres d'intérêt de l'utilisateur. On se base sur l'hypothèse qu'un utilisateur a divers centres d'intérêt et qu'il peut basculer d'un centre d'intérêt vers un autre au cours de sessions de recherches successives. Ainsi, on propose de comparer le contexte courant *cc* avec le contexte extrait à la période précédente *pc*. Nous adoptons, à cet effet une méthode statistique qui scrute le changement dans les centres d'intérêt à court terme [176], en utilisant le coefficient de corrélation des rangs de Kendall comme suit :

$$\Delta C = \frac{\sum_t \sum_{t'} S_{tt'}(pc) * S_{tt'}(cc)}{\sqrt{(\sum_t \sum_{t'} S_{tt'}(pc)^2)(\sum_t * \sum_{t'} S_{tt'}(cc)^2)}} \tag{5.6}$$

où

$$S_{tt'}(pc) = Signe(pc(t) - pc(t')) = \frac{pc(t)-pc(t')}{|pc(t)-pc(t')|},$$
$$S_{tt'}(cc) = Signe(cc(t) - cc(t')).$$

La valeur du coefficient ΔC est dans l'intervalle $[-1..1]$, où une valeur proche de -1 signifie que les contextes sont non similaires, alors qu'une valeur proche de 1 signifie que les contextes sont similaires.

Le coefficient ΔC est approché par une loi Laplace Gauss [154] : $\Delta C \approx LG\left(0; \sqrt{\frac{2(2n+5)}{9n(n-1)}}\right)$ avec n, le nombre de termes traités, soit dans notre cas, $T(R_u^s)$. Le seuil critique de corrélation σ est alors donné par la table correspondante.

L'évolution des centres d'intérêt, qui a pour effet la mise à jour éventuelle de la librairie C^s, est alors déterminée par le résultat de comparaison de ΔC relativement au seuil de corrélation σ. Plus précisément, la stratégie est la suivante :

1. $\Delta C > \sigma$. Les sessions de recherches sont inscrites dans le même contexte : pas d'indication sur l'évolution des centres d'intérêt de l'utilisateur $C^{s+1} = C^s$;
2. $\Delta C < \sigma$. Détection d'un changement de contexte ; deux configurations se présentent : découverte d'un nouveau centre d'intérêt ou évolution d'un autre préalablement découvert. On procède alors de la manière suivante :
   - sélectionner $c^* = argmax_{c \in C^s}(c \circ cc)$, tel que $\circ$ est le coefficient de corrélation ;
   - si $cc \circ c^* > \sigma$ alors :
     - affiner le descripteur du centre d'intérêt $c^* = \delta * c^* + (1-\delta) * cc$, tel que δ est un facteur d'atténuation du poids initial des termes de c^* et renforcer les poids du contexte courant,
     - mettre à jour la matrice H^s par élimination des lignes les moins récemment recalculées R_u^s,
   - si $cc \circ c^* < \sigma$ alors :
     - élargir la librairie des centres d'intérêt c'est-à-dire $C^{s+1} = C^s \cup c^*$,
     - réinitialiser la matrice H^s de manière à privilégier l'apprentissage de ce nouveau centre d'intérêt, poser $s_0 = s$.

Les centres d'intérêt ainsi construits peuvent être réutilisés dans différentes étapes d'un processus d'accès personnalisé à l'information. La librairie C^s constitue alors une ressource pour la définition d'heuristiques et/ou techniques pour la personnalisation de l'accès à l'information. Elle peut ainsi être exploitée dans différentes approches de personnalisation telles que la reformulation de requête, l'adaptation de la fonction d'évaluation de la pertinence,

le réordonnancement des résultats de recherche, etc. Pour notre part, on propose d'exploiter la librairie des centres d'intérêt dans un modèle d'accès personnalisé spécifique que nous présentons dans le chapitre 6.

5.6 Approche implicite pour l'acquisition des données utilisateur

La construction de la dimension « *historique des interactions* » nécessite l'acquisition des documents jugés pertinents par l'utilisateur lors de ses différentes sessions de recherche. Ces jugements peuvent être donnés explicitement par l'utilisateur dans le cas idéal. En pratique, il est rare que l'utilisateur juge ces documents, mais on peut évaluer l'intérêt de l'utilisateur vis-à-vis du document implicitement en observant son comportement lors de sa recherche. Cet intérêt donne une indication sur la pertinence du document vis-à-vis du besoin de l'utilisateur. Ce degré d'intérêt est considéré comme un facteur de discrimination lié à la décision de collecter ou pas le document.

Le processus d'acquisition implicite des documents nécessite en premier la détermination des comportements observables, ainsi que les indicateurs associés. Nous avons préalablement exposé ces différents éléments dans le chapitre 3. Le détail de l'ensemble des comportements recensés par la littérature est présenté dans l'annexe A. Cette typologie, nous permet d'en retenir les plus significatifs.

On propose d'appliquer une démarche permettant de calculer le degré d'intérêt implicitement de l'utilisateur pour un document en combinant les valeurs des indicateurs d'intérêts associés à chaque comportement utilisateur. Cette approche est mise en œuvre à travers l'outil $\mathcal{W}eb\,Cap$ correspondant à l'interface du SRI. La description générale de cet outil est présentée en annexe B. Dans ce qui suit, nous présentons notre démarche d'acquisition implicite de l'intérêt d'un document ainsi que l'évaluation expérimentale de notre outil.

5.6.1 La catégorie de comportement et indicateurs associés

A partir de l'état de l'art sur les indicateurs d'intérêts implicite on retient les comportements suivants : *lecture, sauvegarde, impression*.

1. *La lecture* (*L*)

   C'est un comportement trivial, considéré comme le premier observable lors de l'activité de recherche de l'utilisateur. En effet, si un document a été lu, ceci indique qu'il a suscité un certain intérêt chez l'utilisateur [169]. Ainsi, la connaissance de l'action de lecture d'un document permet de se prononcer sur la susceptibilité de sa pertinence.

2. *La sauvegarde* (*S*)

   La sauvegarde d'un document est un comportement de faible fiabilité pour indiquer l'intérêt de l'utilisateur s'il est considéré comme une action isolée : on peut enregistrer un document par faute de temps. Néanmoins, en le combinant avec les deux autres comportements (lecture et impression), il s'avère être un bon indicateur pour renforcer le degré d'intérêt à inférer.

3. *L'impression* (*I*) considérée isolément peut ne pas traduire un intérêt réel de l'utilisateur. De manière analogue au comportement de sauvegarde, l'impression sera combinée avec les deux autres comportements. A titre d'exemple, si un document est lu et ensuite imprimé, ceci traduit l'intérêt suscité pour l'utilisateur.

L'observation de ces trois comportements s'effectue à l'aide d'indicateurs implicites. Parmi l'ensemble des indicateurs existants, on retient les suivants, car se sont ceux considérés comment étant les plus importants et permettent d'observer au mieux le comportement de l'utilisateur lors de ses différentes sessions de recherches dans la littérature (voir Annexe A) :

1. Indicateurs du comportement « lecture » :

   (a) *La durée de lecture* : est considérée comme le plus important des indicateurs du comportement de lecture. Le temps effectif que prend l'utilisateur pour lire les documents dépend fortement de la longueur des documents. Dans le sens, où plus le document est long, plus la valeur de cet indicateur est grande. Cependant, si cette valeur est élevée pour les documents longs, elle ne traduit pas forcément un intérêt plus grand par rapport aux documents de plus petite taille. Ainsi, on propose d'utiliser une durée de lecture normalisée par rapport à la taille du document lu.

   (b) *Le nombre de clicks de la souris* : plus le nombre est grand, plus le document est susceptible d'intéresser l'utilisateur.

   (c) *Les mouvements de la souris* : permettent de détecter la présence d'une activité sur une page donnée. Le mouvement de la souris se mesure par la distance parcourue de la souris sur l'écran de l'utilisateur. La mesure de distance appliquée dans notre cas est celle proposée par [93], donnée par la formule suivante :

   $$dist\_mouv\_souris(pixels) = \sum_{i=0}^{t-1} Dist(P(t_i) - P(t_i - t_{i-1})) \tag{5.7}$$

   Où : *Dist* est la distance euclidienne, t représente la durée d'activation de la fenêtre, $p(t_i)$ la position de la souris par rapport à l'axe des X et des Y. L'intervalle de temps $t_i - t_{i-1}$ est de 100 ms.

   (d) *Les mouvements du scroll bar*. L'utilité de cet indicateur est que lorsque l'utilisateur juge qu'un document est pertinent, il sera intéressé par le lire entièrement et ceux qu'en visualisant son contenu en le fessant défiler en utilisant la souris et/ou les différentes touches de direction du clavier $(\uparrow, \downarrow, \leftarrow, \rightarrow)$.

2. Indicateurs du comportement « sauvegarde » : il n'existe qu'un seul indicateur à savoir *l'action de sauvegarde*.
3. Indicateurs du comportement « impression » : l'indicateur existant est *la commande d'impression*.

5.6.2 Calcul du degré d'intérêt implicite

Notre outil permet de calculer le degré d'intérêt de l'utilisateur pour chaque document retourné pendant une session de recherche en observant son comportement, puis d'en retenir ceux dont le degré d'intérêt dépasse un certain seuil (fixé par des heuristiques validées expérimentalement).

Plus particulièrement, pour chaque comportement (lecture (L), sauvegarde (S), impression (I)), nous calculons un degré d'intérêt. Ce degré est décliné à l'aide des indicateurs implicites cités précédemment, associés à chacun de ces comportements. Ainsi, lors des interactions de l'utilisateur avec le SRI durant une session de recherche, nous calculons les valeurs effectivement obtenues à travers les actions de l'utilisateur. Si la valeur obtenue par l'indicateur dépasse un seuil fixé expérimentalement, alors nous attribuons à cet indicateur un poids permettant de traduire le degré d'observation du comportement associé. Il en va ainsi pour tous les indicateurs de chaque comportement. A la fin, le degré d'intérêt implicite affecté au comportement sera égal à la somme des poids affectés à chacun des indicateurs réellement observés.

Pour réaliser ce travail, nous avons procédé comme suit :

Les actions possibles réalisées par l'utilisateur peuvent s'effectuer sur la base des hypothèses suivantes :

– *Si la valeur d'un indicateur i associé à la lecture dépasse un seuil prédéfini, son degré d'intérêt sera* α_{L_i} *, à* 0 *sinon.*
– *Si le document est sauvegardé, alors le degré d'intérêt associé* α_S *sera égal à* α_{S_1}*, à* 0 *sinon.*
– *Si le document est imprimé, alors le degré d'intérêt associé* α_I *sera égal à* α_{I_1}*, à* 0 *sinon.*

Ainsi, pour chaque comportement on calculera trois degrés d'intérêt :

* (α_L) : Pour le comportement *Lecture*, nous évaluons les indicateurs permettant de l'observer (L_1, L_2, L_3, L_4) respectivement : la durée de lecture, le nombre de clicks et le mouvement de la souris et enfin le mouvement du scroll bar. Le degré d'intérêt du document par rapport à la lecture noté α_L sera alors la somme des degrés d'intérêts retournés par les quatre indicateurs.
* (α_S) : Pour la sauvegarde, en se basant sur l'indicateur de sauvegarde.
* (α_I) : pour l'impression, en se basant sur l'indicateur d'impression.

L'algorithme suivant résume ces étapes.

Début
Calcul du degré d'intérêt pour les documents de S^s
Pour chaque $d_j \in R_u^s$
Le degré d'intérêt pour la lecture : $(\alpha_L) = \sum_i^4 \alpha_{L_i}$;
(α_S) = degré d'intérêt pour la sauvegarde ;
(α_I) = degré d'intérêt pour l'impression ;
$\alpha_{d_j} = f((\alpha_L),(\alpha_S),(\alpha_I))$
Fin Pour

Le degré d'intérêt (α_{d_j}) associé à un document d_j sera alors fonction des degrés d'intérêts relatifs aux différents comportements observés pour ce document :

$$\alpha_{d_j} = \frac{\sum((\alpha_L)+(\alpha_S)+(\alpha_I))}{3}$$

On considère qu'un document est pertinent si $\alpha_{d_j} \succ Seuilglobal$.

5.6.3 Initialisation des poids des indicateurs implicites

Le degré d'intérêt implicitement associé à un document est obtenu en sommant les degrés d'intérêts obtenus par les comportements de l'utilisateur observés. Ainsi, concernant la sauvegarde et l'impression les degrés assignés, respectivement α_S et α_I, sont des constantes à déterminer. On pose la première hypothèse suivante :

– **Hypothèse 1.** Si un document a été sauvegardé ou bien imprimé, c'est qu'il a suscité l'intérêt de l'utilisateur, donc il doit être considéré comme pertinent.

De plus, étant donné que l'indicateur de lecture est fonction des quatre actions : *durée de lecture*, *nombre de clicks de la souris*, *mouvement de la souris* et *mouvement du scroll bar*, sa quantification dépend des valeurs obtenues par ces actions. Cependant, les valeurs de ces indicateurs ne sont pas sur une même échelle. Comme nous l'avons déjà mentionné, nous assignons à chaque indicateur un poids selon les règles suivantes :

– **Règle 1.**
Si *durée de lecture* $\succ$ *seuil durée lecture* alors $\alpha_{L_1} = poids_{L_1}$, sinon $\alpha_{L_1} = 0$;
– **Règle 2.**
Si *nombre de clicks de la souris* $\succ$ *seuil clicks de la souris* alors $\alpha_{L_2} = poids_{L_2}$, sinon $\alpha_{L_2} = 0$;
– **Règle 3.**
Si *mouvement de la souris* $\succ$ *seuil mouvement de la souris* alors $\alpha_{L_3} = poids_{L_3}$, sinon $\alpha_{L_3} = 0$;
– **Règle 4.**
Si *mouvement du scroll bar* $\succ$ *seuil mouvement du scroll bar* alors $\alpha_{L_4} = poids_{L_4}$, sinon $\alpha_{L_4} = 0$;

Afin de quantifier ces différentes valeurs, nous posons la seconde l'hypothèse suivante :

– **Hypothèse 2.** Un document peut susciter un grand intérêt chez l'utilisateur sans qu'il ne le sauvegarde ou qu'il ne l'imprime. Dans ce cas, le comportement de lecture doit nous permettre de capturer cet intérêt. Comme la durée de lecture est l'indicateur le plus important qui permet d'observer ce comportement, il est nécessaire que son poids soit supérieur à ceux des trois autres indicateurs (nombre de clicks, mouvement de la souris et mouvement du scroll). Mais tout seul, cet indicateur ne traduit pas vraiment l'intérêt de l'utilisateur, car il ne prouve pas la présence réelle de l'utilisateur. D'où la nécessité de l'associer à l'un des deux autres. Si la durée de lecture et au moins l'un des deux indicateurs, activité de la souris et mouvement du scroll, ont dépassé leurs seuils respectifs le document doit être jugé pertinent.

Le choix des différents poids des indicateurs dépend donc des deux hypothèses ci-dessus ainsi que du seuil global.

La première hypothèse est traduite formellement comme suit :

Le document est pertinent ssi $\alpha_S > seuilglobal$ ou $\alpha_I > seuilglobal$.

La seconde hypothèse est traduite formellement comme suit :

Le document est pertinent ssi $\sum_i^4 \alpha_{L_i} > seuilglobal \wedge \alpha_{L_1} > \alpha_{L_2}, \alpha_{L_3}, \alpha_{L_4}$.

Afin d'effectuer nos expérimentations nous avons fixé de manière arbitraire seuil global à **0.3**. Nous avons également considéré pour les différents poids des indicateurs les valeurs suivantes :

– Un poids égal à **0.3** pour chacun des comportements sauvegarde et impression ($S_1 = 0.3, I_1 = 0.3$)
– Un poids maximal de **0.4** est accordé au comportement lecture réparti comme suit :
    * Pour l'indicateur « *durée de lecture* » : 0.2 ;
    * Pour l'indicateur « *nombre de clicks* » : 0.05 ;
    * Pour l'indicateur « *mouvement de la souris* » : 0.05 ;
    * Pour l'indicateur « *mouvement du scroll* » : 0.1 ;

Concernant les valeurs des seuils accordés aux différents indicateurs α_{L_i} du comportement *lecture*, elles sont déterminées expérimentalement lors de la phase d'initialisation de notre outil.

5.6.4 Validation expérimentale

Afin d'évaluer l'impact de notre approche à inférer l'intérêt de l'utilisateur pour un document, nous avons construit un cadre expérimental pour comparer les degrés d'intérêt explicitement donnés par l'utilisateur et ceux calculés implicitement.

La campagne s'est effectuée avec six utilisateurs, étudiants en informatique. A chaque session de recherche (soumission d'une requête), les utilisateurs doivent juger explicitement chaque document qu'ils consultent. Pour ces mêmes documents on calcule le degré d'intérêt

implicitement à l'aide de l'outil $\mathcal{W}eb\,Cap$. Le jugement explicite donné par l'utilisateur est obtenu à l'aide d'une fenêtre de l'outil $\mathcal{W}eb\,Cap$. Le jugement implicite est obtenu automatiquement par la démarche implémentée dans $\mathcal{W}eb\,Cap$.

Les expérimentations ont duré cinq jours à raison d'une session de recherche par jour pour chaque utilisateur. La durée de chaque session de recherche a été fixée à environ 1 heure. Les expérimentations ont été menées en deux phases. Dans un premier temps (le premier jour), les résultats nous ont permis d'initialiser les seuils des indicateurs du comportement *lecture* ; par la suite (les quatre derniers jours), nous avons procédé aux évaluations comparatives des résultats. Dans ce qui suit, nous allons détailler les différentes étapes de cette phase d'expérimentation.

5.6.4.1 Phase d'initialisation

Le calcul du degré d'intérêt implicite d'un document est effectué selon la démarche décrite précédemment. Ceci revient à comparer les valeurs capturées par observations des indicateurs aux seuils associés. Il est donc indispensable dans un premier temps, de fixer les valeurs des seuils des indicateurs d'intérêt implicites, afin de commencer à évaluer la capacité de l'approche implicite et de les modifier par la suite durant la seconde phase pour aboutir à un meilleur rendement.

Pour initialiser ces valeurs, nous avons procédé comme suit : le premier jour de l'expérimentation et pour chaque utilisateur effectuant une recherche, nous avons collecté pour chaque document consulté les valeurs des indicateurs suivants : la durée de lecture, le nombre de clicks, le mouvement de la souris et le mouvement du scroll bar.

Pour chaque indicateur, nous avons calculé un seuil en prenant la moyenne des valeurs obtenues par les six utilisateurs pour l'ensemble des documents consultés selon la formule suivante :

$$Seuil\_Indicateur_i = \frac{\sum_{j=1}^{6}\sum_{k=1}^{R}(Valeur\_Observation(Indicateur_i)_{jk})}{6\star R} \tag{5.8}$$

Où, R est le nombre de documents jugés par les six utilisateurs, fixé à dix documents par utilisateur ; $Valeur\_Observation(Indicateur_i)$ est la valeur réelle de la réalisation de l'action utilisateur sur le document correspondant à l'indicateur i.

Ainsi, on résume dans le tableau 5.4 les valeurs des seuils obtenus pour ces indicateurs d'intérêt par les six utilisateurs face à dix documents retournés en appliquant la fonction 5.8.

Seuil indicateur L_1	**Seuil indicateur** L_2	**Seuil indicateur** L_3	**Seuil indicateur** L_4
72 *seconde*	22	45	35

Tableau 5.4 — Seuils calculés pour chaque indicateur

5.6.4.2 Analyse des résultats

Durant cette phase, qui a duré cinq jours, chaque utilisateur a effectué au total cinq requêtes (à raison d'une requête par jour), nous avons collecté les résultats de recherche à savoir les valeurs des différents indicateurs ainsi que les degrés d'intérêt implicites et explicites. A la fin des quatre jours, nous avons collecté les données suivantes pour les six utilisateurs :

* le nombre de documents consultés est de 243,
* le nombre de documents jugés explicitement pertinents est de 138,
* le nombre de documents jugés implicitement pertinents est de 110.

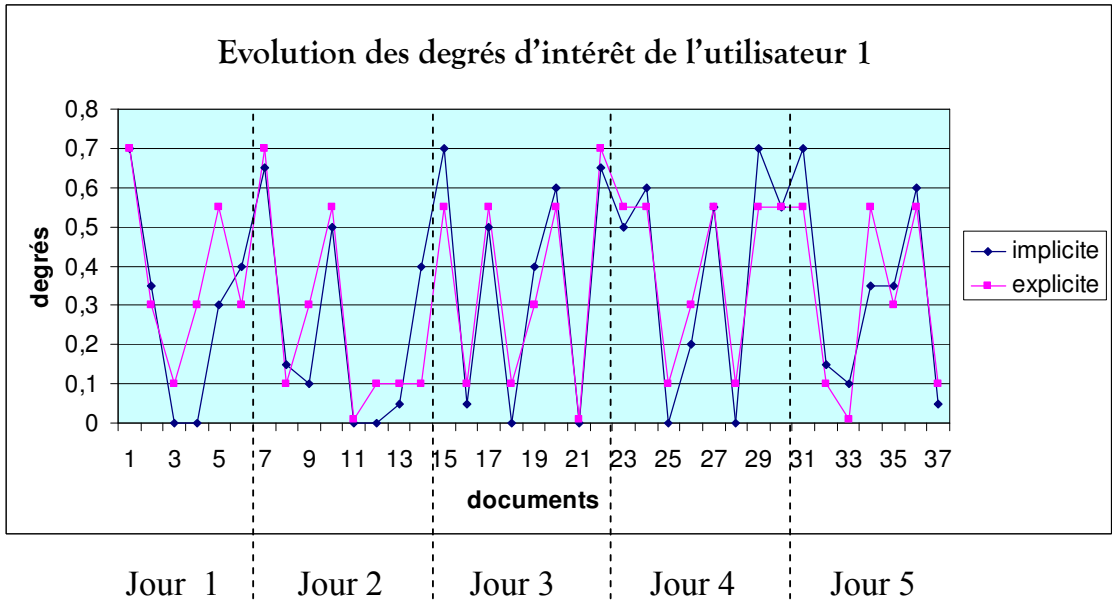
Afin de pouvoir comparer, pour chaque document jugé, son degré de pertinence explicitement donné par l'utilisateur (non pertinent, faiblement pertinent, moyennement pertinent, pertinent, très pertinent) avec celui que l'on calcule implicitement, il est nécessaire de traduire ce jugement de pertinence explicite par des poids qui sont dans une même échelle de valeur que les degrés d'intérêts implicites. Pour cela, nous traduisons les degrés de pertinence en poids, comme suit :

– Non pertinent : 0.01 (pour éviter le 0 qui signifie absence d'un jugement explicite).
– Faiblement pertinent : 0.1.
– Moyennement pertinent : 0.3 (si un document est jugé explicitement moyen il sera sauvegardé dans l'historique des interactions).
– Pertinent : 0.55.
– Très pertinent : $>= 0.7$.

Les graphes des figures 5.2, 5.3 et 5.4 résument l'évolution des degrés d'intérêt implicites et explicites pour chaque utilisateur durant les cinq jours de l'expérimentation.

D'une manière générale, les deux courbes évoluent de la même manière, et sont proches l'une de l'autre, ce qui traduit que le degré d'intérêt implicite correspond à la pertinence effective. On remarque par ailleurs que lorsque les documents jugés explicitement moyens, leurs degré implicite ne correspond pas toujours au degré explicite, on le constate par exemple pour les documents entre 23 et 29 de l'utilisateur 2. Ceci pourrait être expliqué par l'hésitation de l'utilisateur concernant son intérêt réel pour ces documents. En effet, lorsqu'un utilisateur exprime un jugement de pertinence moyen (ni bon, ni mauvais), le degré implicite calculé est généralement faible.

Les résultats obtenus démontrent que l'on arrive à collecter environ 80% (110 sur 138) de documents pertinents de manière implicite, comparativement aux jugements explicitement donnés par les utilisateurs. Les 20% restants sont en général le résultat donné parmi les documents jugés moyens, ce qui peut être expliqué par le fait qu'un utilisateur ne peut pas toujours donner son intérêt exact pour un document, surtout quand celui ci ne suscite pas un très grand intérêt.



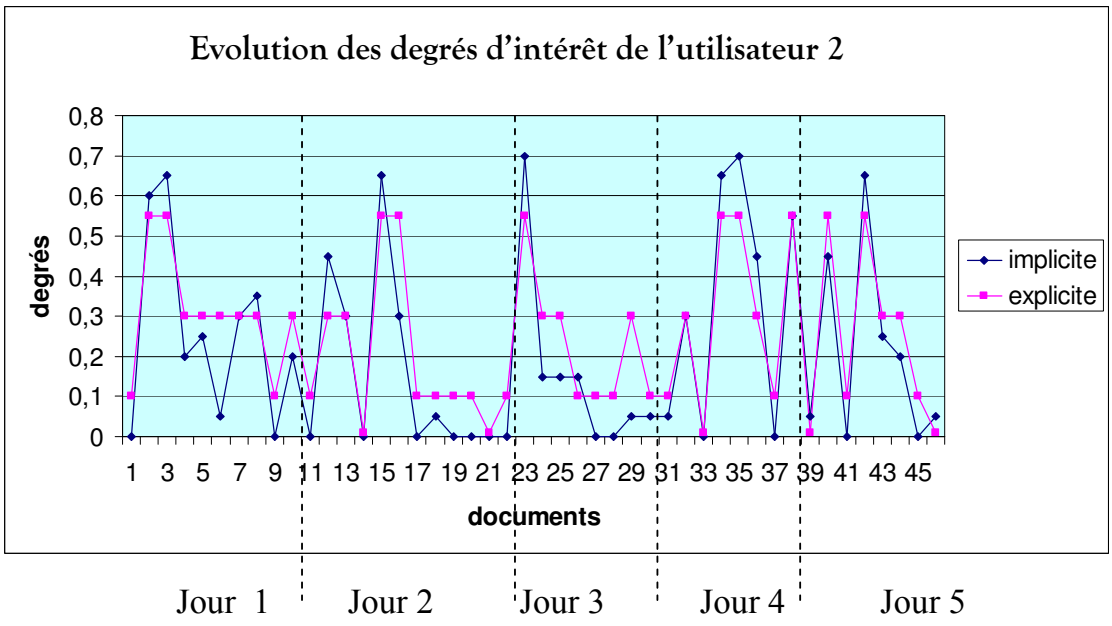
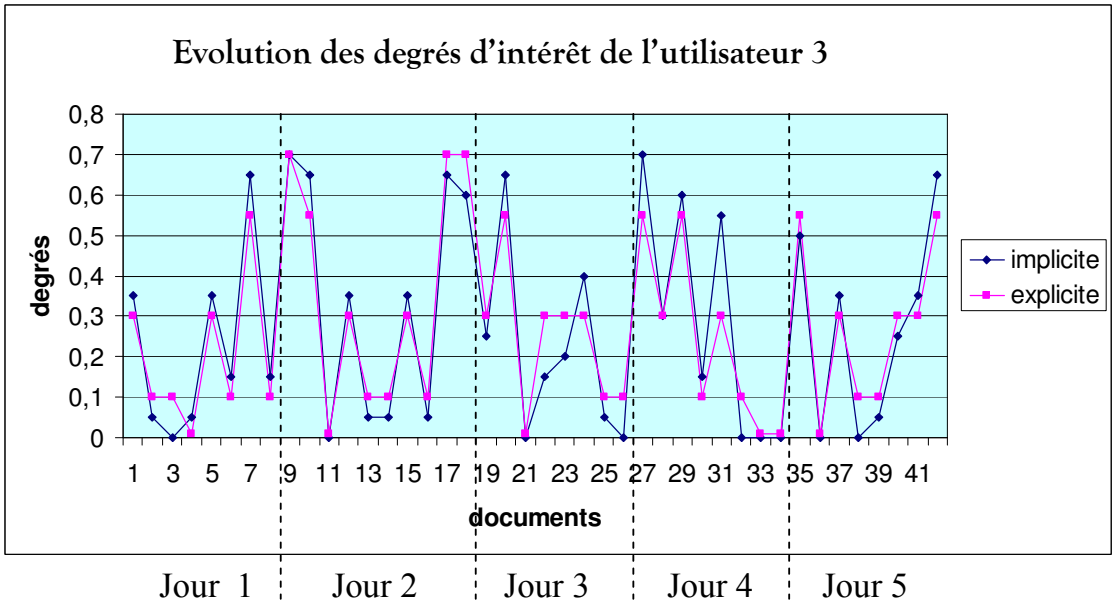


Figure 5.2 — Résultats expérimentaux des utilisateurs 1 & 2



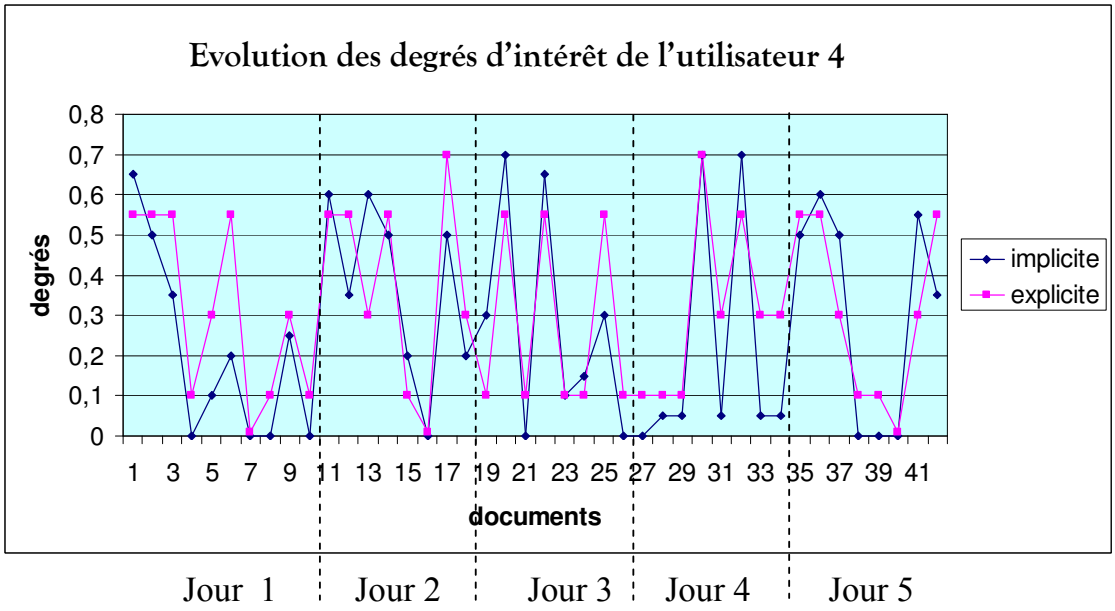
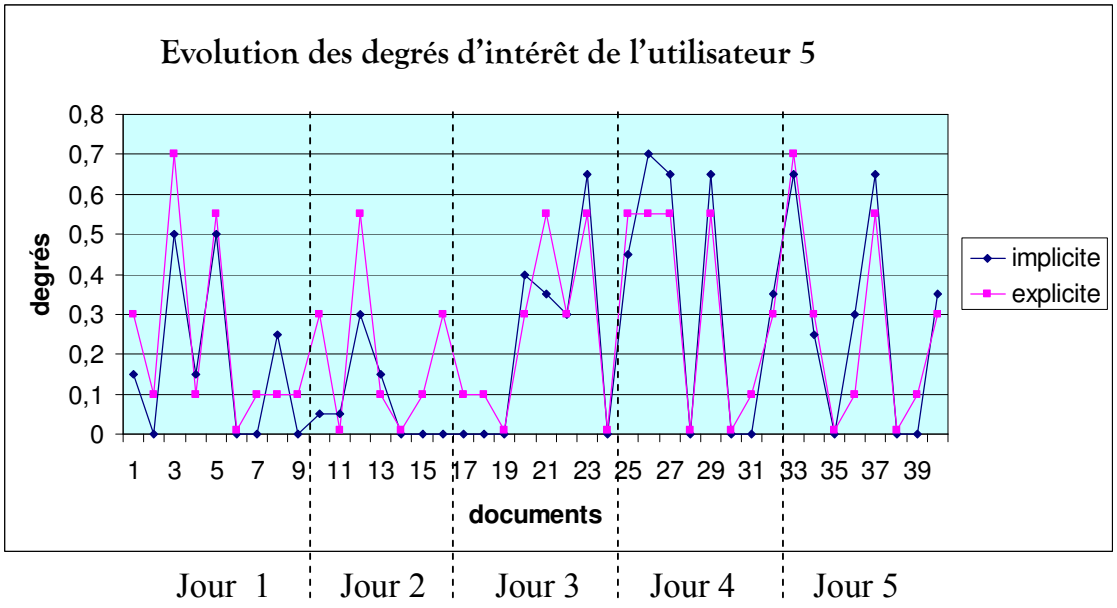


***Figure 5.3* —** Résultats expérimentaux des utilisateurs 3 & 4



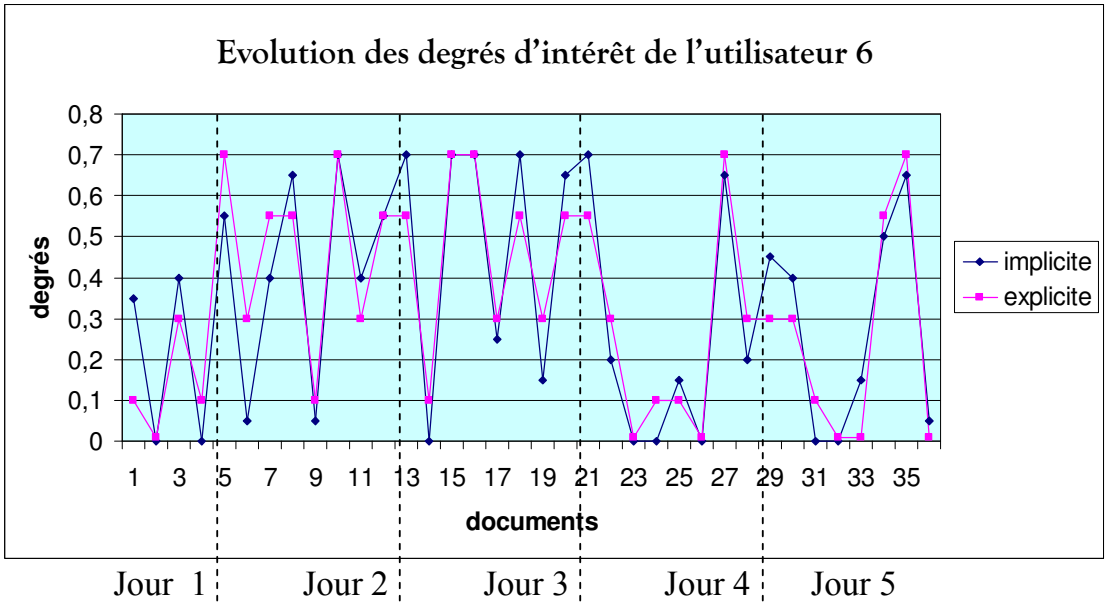


Figure 5.4 — Résultats expérimentaux de l'utilisateur 5 & 6

5.7 Conclusion

Notre contribution pour la modélisation du profil utilisateur porte sur deux volets. Le premier concerne la définition d'un profil utilisateur décrit par deux dimensions informationnelles. La première correspond à l'historique de recherche de l'utilisateur. Elle est construite en agrégeant les informations collectées implicitement/explicitement lors des sessions de recherches successives. La seconde dimension traduit les centres d'intérêt de l'utilisateur dérivés automatiquement à partir de l'historique des interactions. Le profil évolue selon une approche statistique basée d'une part, sur la distribution des termes dans les documents jugés explicitement ou implicitement pertinents et d'autre part, sur une mesure de corrélation permettant de scruter le changement des centres d'intérêt de l'utilisateur au cours du temps.

L'approche de définition du profil ainsi présentée est essentiellement caractérisée par :

1. la définition et utilisation d'une mesure de pertinence relative des termes pour un profil : cette mesure considère l'information véhiculée par les documents jugés par l'utilisateur,
2. l'exploitation de la seule dimension historique des recherches de l'utilisateur pour la construction et évolution de ses centres d'intérêts : aucune autre ressource n'est requise,
3. la prise en compte de l'évolution et la diversité des centres d'intérêt de l'utilisateur au cours du temps. On utilise une méthode statistique pour maintenir la diversité des centres d'intérêts : cette méthode produit, à terme, une librairie qui peut être utilisée comme ressource pour personnaliser l'accès à l'information.

Des points méritent cependant d'être affinés. Le plus important porte sur la définition des périodes d'évolution des centres d'intérêt. En effet, la variation des centres d'intérêt de l'utilisateur, décelée à travers les requêtes qu'il a émis, ne présente pas forcément des régularités prévisibles. Même si ce risque pourrait être amoindri en réduisant au mieux ces périodes, une perspective intéressante est de mener une réflexion plus poussée sur un compromis entre les différents paramètres qui régulent l'évolution des centres d'intérêts d'un utilisateur.

Le second volet de notre contribution abordé dans ce chapitre, est l'application d'une approche d'acquisition implicite des documents visités par l'utilisateur. Ces documents sont collectés par estimation d'un degré d'intérêt implicitement inféré du comportement de l'utilisateur lors de sa session de recherche. Notre travail a abouti au développement d'un outil *Web Cap* implémentant l'approche d'acquisition implicite proposée. L'évaluation expérimentale des performances de notre approche, nous a permis de constater une capacité égale à 80% d'acquisition des documents pertinents de manière implicite comparativement à l'approche explicite.

6 Modèle d'Accès Personnalisé à l'Information basé sur les Diagrammes d'Influence

6.1 Introduction

Ce chapitre est consacré à la définition d'un nouveau modèle d'accès personnalisé à l'information intégrant un profil utilisateur, particulièrement celui proposé dans le chapitre précédent. L'idée de base de notre modèle est de substituer à la fonction de pertinence classique qui mesure le degré d'appariement requête-document $RSV(q,d) = P(d/q)$, une fonction indexée par l'utilisateur $RSVu(q,d) = P(d/q,u)$, où $P(a/b)$ est la probabilité conditionnelle de l'événement A sachant l'événement B et U représente l'utilisateur décrit par ses centres d'intérêts. Ainsi, pour formaliser cette fonction de pertinence, notée $RSVu(q,d)$, on s'oriente vers l'utilisation d'une extension des Réseaux Bayésiens [135; 85] en l'occurrence les diagrammes d'influence [**?**].

Le choix de ce formalisme est motivé par l'objectif de proposer un modèle inférentiel pour la mise en œuvre du raisonnement lié à la prise de décision quant à la pertinence des documents compte tenu du profil de l'utilisateur d'une part et de la requête d'autre part. Le calcul du score de pertinence d'un document est fondé sur une fonction permettant d'estimer l'utilité de présenter à l'utilisateur ce document et ce, compte tenu de sa requête et de ses centres d'intérêts spécifiques. Les Réseaux Bayésiens apportent donc des outils permettant de traiter l'incertitude à laquelle nous sommes confrontés dans le cadre de l'intégration du profil utilisateur dans le processus de RI, et qui réside dans :

- La complexité liée à la représentation des entités intervenant dans le processus de RI (requête, documents, profil utilisateur) et des relations entre eux ;
- L'évaluation de la pertinence des documents vis-à-vis de la requête et du profil utilisateur.

Preuve en est, l'utilisation de ce formalisme par de nombreux auteurs, qui ont défini plusieurs manières de considérer les relations de dépendance entre documents et requête, d'observer les documents et de satisfaire le besoin utilisateur. Ce sont ces mêmes points que nous allons aborder, dans ce chapitre, à travers nos contributions qui se distinguent des approches

précédemment citées par l'intégration du profil de l'utilisateur, plus précisément ses centres d'intérêts, dans la formalisation du processus d'accès à l'information.

Dans la suite de ce chapitre, nous présentons dans la section 6.2 les principes théoriques des réseaux Bayésiens et des Diagrammes d'influence (DI). Dans la section 6.3, nous spécifions formellement l'objectif du modèle proposé : mettre en œuvre un appariement *requête-document-utilisateur*. Dans la section 6.4, nous décrivons l'architecture générale de notre modèle représentant la composante qualitative du modèle. Dans la section 6.5, nous présentons le processus d'évaluation de la requête. L'estimation des distributions des probabilités attachées aux variables du modèle d'accès, représentant la composante quantitative du modèle, est abordée dans la section 6.6. Dans la section 6.7, nous discuterons du choix de l'opérateur d'agrégation utilisé lors du processus d'évaluation de la requête. Nous concluons ce travail dans la section 6.8.

6.2 Cadre formel

La modélisation graphique telle que les Réseaux Bayésiens, les diagrammes d'influence, les chaînes de Markov, les arbres de décision, etc., s'appuient sur la théorie des probabilités. Les modèles probabilistes constituent des outils puissants pour la RI car ils permettent de traiter d'une manière efficace l'incertitude intrinsèque au processus, principalement dans la décision de considérer un document comme pertinent. En effet, plusieurs travaux ont tenté d'exploiter l'apport des Réseaux Bayésiens (RBs) pour définir des modèles de RI [135; 183; 184; 48; 32].

6.2.1 Les Réseaux Bayésiens

Les réseaux Bayésiens (RB) sont des modèles graphiques capables de représenter et manipuler efficacement des distributions de probabilités multidimensionnelles. Un RB utilise deux composantes pour codifier une connaissance qualitative et quantitative. La composante qualitative est représentée par un graphe acyclique $G = (V,E)$ où V comprend des nœuds qui représentent des variables aléatoires $(X_1,\ X_2,\ldots,\ X_n)$ et E un ensemble d'arcs entre nœuds qui traduisent des relations de causalité. A chaque variable X_i est associée un ensemble de valeurs mutuellement exclusives définies dans $dom(X_i)$. La composante quantitative comprend des distributions de probabilités conditionnelles où pour chaque variable $X_i \in V$ est attachée une classe de probabilités $P(X_i/pa(X_i))$ avec $pa(X_i)$ une combinaison de valeurs associées aux parents de X_i dans G. Un RB permet une représentation compacte de la loi jointe :

$$P(X) = \prod_{i=1}^{n} P(X_i/pa(X_i))$$

Le principal avantage apporté par l'utilisation des RBs en RI a été de pouvoir combiner des informations provenant de différentes entités (requêtes, termes, documents) pour resti-

tuer les documents qui seraient les plus pertinents étant donnée une requête. D'une part, la composante qualitative du réseau permet de représenter par des nœuds, les documents de la collection, les termes d'indexation, la requête et l'utilisateur, et de représenter les relations de dépendance (ou d'indépendance) existantes entre ces variables par des arcs. L'aspect quantitatif du réseau permet, quant à lui, d'évaluer les arcs reliant toute paire de nœuds sur la base de calcul de probabilités. Ainsi, un processus d'inférence permet d'attribuer des degrés de croyance à de nouveaux événements.

La première utilisation des RBs en RI est apparue dans les années 80 [63] mais s'est largement développée par les travaux de Turtle [183; 184]. Turtle et Croft [183; 184] ont proposé un modèle de RI basé sur les réseaux d'inférence Bayésiens pour calculer la probabilité de pertinence d'un document étant donnée une requête. Ce modèle a été étendu pour inclure la réinjection de pertinence (*relevance feedback*). D'autre modèle de RI basé sur les réseaux possibilistes ont été proposés par [33; 22]. Nous rappelons en annexe C quelques notions de RBs ainsi que les méthodes de calcul des probabilités conditionnelles dans ces réseaux.

6.2.2 Les Diagrammes d'Influence

Dans les RBs, les décisions ne sont pas explicitement modélisées, c'est pourquoi les arbres de décision et les diagrammes d'influence ont été développés, pour permettre une meilleure modélisation du processus de prise de décision [**?** 85]. Ces formalismes sont basés sur les RBs, mais contiennent deux types de nœuds supplémentaires :

– Les nœuds de décision : représentent la décision à prendre. Ces nœuds ne possèdent pas de table de probabilité ;
– Les nœuds d'utilité : sont utilisés pour fournir une représentation quantitative des effets des décisions. Ces nœuds n'ont ni état ni table de probabilité et ne peuvent avoir de nœuds-fils. Un nœud d'utilité contient une valeur numérique (utilité) pour chaque combinaison d'états de ses nœuds-pères, cette valeur peut représenter un coût, une durée ou n'importe quelle autre mesure de performance, le but du DI étant de maximiser (ou minimiser) la valeur d'utilité.

Généralement, un RB peut être étendu en diagramme d'influence [167] :

* Les nœuds de décision sont ajoutés là où une décision doit être prise ou un test effectué ;
* Les parents des nœuds de décision sont les variables qui affecteront le résultat de la décision ;
* Les enfants des nœuds de décision seront, eux, affectés par la décision ;
* Les nœuds d'utilité sont introduits partout où la valeur d'utilité peut être mesurée et sont, généralement, en bas du réseau, permettant ainsi la prise en compte de tous les facteurs.

6.3 Spécification du modèle d'accès personnalisé à l'information

Intuitivement, la problématique de la personnalisation en RI peut être exprimée fondamentalement comme suit [173] :

Etant donné une requête q, l'objectif du SRI est d'identifier les documents d qui sont appropriés au besoin en information de l'utilisateur u.

D'un point de vue probabiliste, cet objectif peut être formulé ainsi :

Etant donné une requête q, l'objectif du SRI est de retrouver les documents susceptibles de répondre à ce besoin pour lesquels la probabilité de pertinence du document d, considérant la requête q et l'utilisateur u, noté $P(d/q,u)$, est la plus élevée.

Cette probabilité est formulée en appliquant la loi de Bayes comme suit :

$$P(d/q,u) = \frac{P(q/d,u)P(d/u)}{P(q/u)} \tag{6.1}$$

où d, q et u sont les variables aléatoires associées à D, Q et U respectivement.

Etant donné que le dénominateur $P(q/u)$ est constant pour une requête et un utilisateur donné, nous pouvons tenir compte uniquement du numérateur afin d'ordonner les documents selon leur pertinence. Ainsi la formule de *RSV* (*Relevance Status Value*) d'un document peut être définie comme suit :

$$RSV_u(q,d) = P(q/d,u)P(d/u) \tag{6.2}$$

Cette expression de la pertinence d'un document, vis-à-vis de la requête et des centres d'intérêts de l'utilisateur, permet de quantifier en termes probabiliste les dépendances entre *requête*, *document* et *utilisateur*. En effet, le premier membre de cette équation ($P(q/d,u)$) est une probabilité conditionnelle associée à la requête q. Il traduit le degré de satisfaction de la requête en tenant compte du document d et l'utilisateur u. Le second membre, indépendant de la requête, exprime en revanche la probabilité de pertinence du document d en tenant compte de l'utilisateur u.

Ainsi, si on considère que l'utilisateur U peut avoir des centres d'intérêts multiples de nombre n, modélisés par $C_1, C_2, ..., C_n$, alors l'équation 6.2 du *RSV* devient :

$$RSV_u(q,d) = P(q/d,c_1,c_2,...,c_n)P(d/c_1,c_2,...,c_n) \tag{6.3}$$

où c_k représente la variable aléatoire associée au centre d'intérêt C_k de l'utilisateur.

On constate d'après cette formulation deux points importants :

1. Deux conditions sont nécessaires quant à la décision de sélection d'un document :

   (1) *pertinence thématique* exprimant l'adéquation du contenu du document au besoin formulé par la requête ;

(2) *condition d'utilité* exprimant le degré d'adéquation du document D à l'ensemble des centres d'intérêts de l'utilisateur.

2. La couverture thématique d'un document au besoin en information désiré par l'utilisateur est réalisée en maximisant la corrélation de l'information aux différents centres d'intérêts de l'utilisateur. Ainsi, on exprime le degré de pertinence nécessaire pour intégrer tout ou un sous ensemble des centres d'intérêts de l'utilisateur durant le processus de personnalisation.

Dès lors, notre objectif est de tenir compte d'un profil utilisateur multi-centres d'intérêts. Cette problématique peut être exprimée globalement à travers le diagramme d'influence, noté $ID(D,C,\mu)$, contenant :

– L'ensemble des variables $D = \{d_1, d_2, ..., d_m\}$ représentant les documents , où n est le nombre total de document dans la collection ;
– L'ensemble des variables $C = \{c_1, c_2, ..., c_n\}$ représentant les centres d'intérêts de l'utilisateur, où c_n est le n^{ime} centre d'intérêts considéré de la librairie des centres d'intérêts de son profil ;
– L'ensemble des utilités $\mu = \{\mu_1, \mu_2, ..., \mu_n\}$, où μ_k exprime l'utilité du document instancié, D pour le centre d'intérêt C_k de l'utilisateur.

Ainsi, l'objectif revient alors à ordonner les documents $d_j \in D$ selon la valeur de l'utilité exprimée par $\mu(d_j) = \Psi(\mu_1, \mu_2, ..., \mu_n)$, où Ψ est un opérateur approprié d'agrégation combinant les valeurs des évidences issues de l'ensemble $c_1, c_2, ..., c_n$. En considérant l'équation 6.2, le calcul du score de pertinence est exprimé par :

$$RSV_U(q,d) = \Psi_{k=1..n}(\mu_k(d,c_k) * p(q/d, c_1, c_2, ..., c_n)) \tag{6.4}$$

La section suivante donne les détails de notre modèle d'accès personnalisé basé sur les spécifications décrites ci-dessus.

6.4 Architecture générale du modèle

La topologie du modèle d'accès est représentée dans la figure 6.1 par le graphe acyclique $G = (V,E)$ où V comprend les nœuds représentant des variables aléatoires $X_1,\ X_2,\ \ldots,\ X_n$. A chaque variable X_i est associée un ensemble de valeurs mutuellement exclusives définies dans $dom(X_i)$. L'ensemble E comprend les arcs existants entre les nœuds qui traduisent des relations de causalité décrites par des probabilités conditionnelles attachées à chaque nœud.

6.4.1 Description des nœuds

D'un point de vue qualitatif, les nœuds composant le graphe représentent différents types d'information exprimés par trois sous-ensembles de nœuds, décrits comme suit :

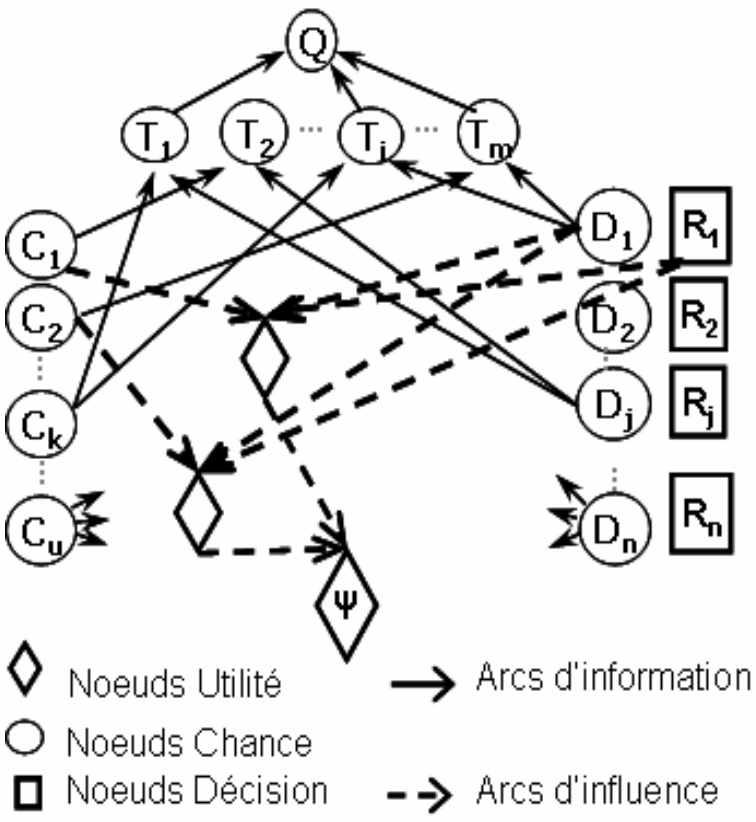


Figure 6.1 — Topologie du modèle d'accès personnalisé à l'information

6.4.1.1 Nœuds *chance*

Ils représentent l'ensemble de variables aléatoires binaires utilisées dans notre modèle exprimées par l'ensemble $V^{info} = Q \cup D \cup T \cup C$.

On distingue quatre types de nœud *chance* :

- Le nœud unique Q correspondant à *la requête* de l'utilisateur est représenté par une variable aléatoire binaire définie dans l'ensemble $dom(Q) = \{q, \overline{q}\}$, où q désigne que la requête Q est *satisfaite* et $\overline{q}$ désigne *que la requête n'est pas satisfaite* ; dans notre cas, on ne s'intéressera qu'à l'instanciation positive q. ($dom(Q)$ désigne l'ensemble de valeurs possibles pour la variable Q)
- L'ensemble $D = \{D_1,\ D_2,\ ...,\ D_m\}$ correspond aux *documents* de la collection. Chaque noeud document D_j, représente une variable aléatoire binaire prenant des valeurs dans l'ensemble $dom(D_j) = \{d_j, \overline{d_j}\}$, où d_j désigne que *le document D_j est observé*, et $\overline{d}_j$ désigne que *le document D_j n'est pas observé*. Un document *observé*, selon l'approche de Turtle [183], traduit un document tiré aléatoirement de la collection et dont on calcule le degré de pertinence.
- L'ensemble $C = \{C_1,\ C_2,\ ...,\ C_n\}$ correspond aux *centres d'intérêts* associés à l'utilisateur. Chaque noeud centre d'intérêt C_k représente une variable aléatoire binaire prenant des valeurs dans l'ensemble $dom(C_k) = \{c_k, \overline{c_k}\}$,où c_k désigne que *le centre d'intérêt C_k*

est observé et $\overline{c_k}$ désigne que *le centre d'intérêt C_k n'est pas observé*. Un centre d'intérêt *observé*, est par analogie au document, un centre tiré aléatoirement du profil de l'utilisateur et dont on calcule le degré de pertinence relativement à la requête. La pertinence d'un centre d'intérêt traduit le fait qu'il couvre l'objet de la requête.

– L'ensemble $T = \{T_1,\ T_2,\ \ldots,\ T_l\}$ correspond aux termes d'indexation. Chaque noeud terme T_i représente une variable aléatoire binaire prenant des valeurs dans l'ensemble $dom(T_i) = \{t_i, \overline{t_i}\}$, où t_i désigne que *le terme T_i est pertinent pour la requête Q* et $\overline{t_i}$ désigne que *le terme T_i n'est pas pertinent pour Q*. La pertinence d'un terme signifie sa présence éventuelle dans un document observé.

6.4.1.2 Nœuds *décision*

On associe à chaque document D_j de la collection, un noeud décision R_j prenant ses valeurs dans $dom(R_j) = \left\{r_j, \overline{r_j}\right\}$ ce qui correspond respectivement aux décisions de présenter ou pas le document D_j à l'utilisateur et ce, compte tenu de sa requête et de son profil décrit par ses centres d'intérêts.

6.4.1.3 Nœuds *utilité*

Un *nœud utilité* exprime *l'utilité de la décision de présenter un document compte tenu des centres d'intérêts* de l'utilisateur. De ce fait, on associe un nœud utilité à chaque document D_j et chaque centre d'intérêt C_k, ce nœud correspondant à l'évaluation de la pertinence du document D_j vis-à-vis de la requête, et au regard du centre d'intérêt C_k.

Les valeurs restituées par le sous-ensemble de ces nœuds concernant un document D_j de la collection sont utilisées par un noeud utilité particulier qui les intègre dans le calcul de *l'utilité globale de la décision de restituer ce document D_j en considérant tous les centres d'intérêts* de l'utilisateur.

6.4.2 Description des arcs

On distingue deux types d'arcs : arcs d'information et arcs d'influence.

– **Les arcs d'information**
Ils reflètent la dépendance entre les valeurs d'importance du document, du centre d'intérêt et des termes les indexant. Ils relient chacun des nœuds termes $T_i \in T$ à chaque nœud document $(D_j) \in D$ qu'ils indexent et ainsi qu'à chaque nœud contexte $(C_k) \in C$ qu'ils représentent. Il existe également des arcs d'information qui relient chaque terme T_i avec le nœud requête Q.

On note $pa(.)$ l'ensemble des parents pour chaque nœud du réseau :
$\forall T_i \in T, pa(T_i) = \tau(D_j) \cup \tau(C_k)$, $T_i \in \tau(D_j)$ et $T_i \in \tau(C_k)$
$\forall D_j \in D, pa(D_j) = \oslash$,
$\forall C_k \in C, pa(C_k) = \oslash$.
où $\tau(D_j)$ et $\tau(C_k)$ représentent les termes d'indexation représentant le document et le centre d'intérêt (respectivement).

– **Les arcs d'influence**
Ce type d'arcs traduit le degré d'influence des variables associées à la décision prise. Dans le cas de notre modèle, des arcs d'influence relient servent à traduire deux types de relations liées à la prise de décision :
  – relier les noeuds décisions aux nœuds centres d'intérêts et documents afin de mesurer les utilités élémentaires associées à la décision de présenter le document compte tenu du centre d'intérêt ;
  – relier les utilités élémentaires de chaque nœud document avant l'ensemble des nœuds centre d'intérêt selon un opérateur d'agrégation Ψ. A ce niveau les arcs d'influence permettent de traduire l'utilité globale de la décision de restituer le document observé en considérant tous les centres d'intérêts.

6.5 Evaluation de la requête

Le processus d'évaluation de requête (calcul de la pertinence) revient, dans notre modèle, à instancier chaque document de la collection et chacun des centres d'intérêt et à calculer la croyance de satisfaire la requête étant donné le document instancié, ceci en instanciant également, un à un, chacun des centres d'intérêt de l'utilisateur. Il s'agira donc d'appliquer un algorithme de propagation des nouvelles croyances (document et centre d'intérêt observés) selon la structure du DI et d'arriver à mesurer la pertinence d'un document vis-à-vis de la requête considérée et de l'ensemble des centres d'intérêt de l'utilisateur.

Plus précisément, soit une requête Q, le processus d'évaluation est enclenché comme dans un problème décisionnel [85] en maximisant les utilités globales exprimés par l'équation. Le rang de pertinence de chaque document peut être exprimé selon la fonction suivante :

$$RSV_U : \begin{cases} R \longrightarrow R \\ RSV_U(Q,D) \mapsto \frac{EU(r_j/D)}{EU(\overline{r_j}/D)} \end{cases} \tag{6.5}$$

où $EU(r_j/D)$ (resp. $EU(\overline{r_j}/D)$) est l'utilité globale associée à la décision "*D est pertinent, peut être présenté à l'utilisateur*" (resp.*"D n'est pas pertinent, ne sera pas présenté à l'utilisateur"*).

A partir de l'architecture du réseau, on en déduit les utilités globales formulées comme suit :

$$EU(r_j/D) = \Psi_{k=1..n}\left[\mu_k(r_j/d,c_k) * P(q/d,c_k)\right] \tag{6.6}$$

$$EU(\overline{r_j}/D) = \Psi_{k=1..n}\left[\mu_k(\overline{r_j}/d,c_k) * P(q/d,c_k)\right] \tag{6.7}$$

La valeur d'utilité globale attendue $EU(r_j/D)$, qui représente l'utilité de restituer un document, découle de l'algorithme de propagation dans le réseau. Cette propagation est liée à la structure du diagramme d'influence qui, prise dans sa globalité, représente les dépendances existantes entre une requête et les documents de la collection ainsi que les centres d'intérêts de l'utilisateur.

Un seul document est instancié positivement ($D_j = r_j$) à la fois, il en est de même pour les centres d'intérêts C_k. La propagation de l'information est déclenchée par ces instanciations. La propagation dans ce modèle consiste alors à calculer, pour chaque noeud, la probabilité *a posteriori* étant donné les probabilités conditionnelles et marginales *a priori*. La propagation tente de calculer la probabilité que la requête soit rencontrée étant donné un document instancié à $D_j = r_j$ et un centre d'intérêts instancié à $C_k = c_k$. Pour chaque document, ce processus est réitéré pour tous les centres d'intérêts ; puis l'algorithme reprend pour considérer tous les documents de la collection candidats à l'évaluation.

Dans le cas général la probabilité $P(q/d,c_k)$ est exprimée par la formule suivante :

$$P(q/d,c_k) = \frac{P(q,d,c_k)}{P(d,c_k)} \tag{6.8}$$

Grâce au principe de **marginalisation**, on obtient l'expression suivante :

$$P(q,d,c_k) = \sum_{\theta^s \in \theta} P(q,\theta^s,d,c_k) \tag{6.9}$$

Où θ représente l'ensemble des configurations possibles des termes d'indexation de $pa(Q)$, θ^s la configuration d'ordre s associée, et θ_i^s la configuration d'ordre s associée au terme $t_i \in pa(Q)$

A titre d'exemple, les configurations possibles des termes de la requête, Q, composées des termes $\{T_1,T_2\}$ sont $\theta = \{\{t_1,t_2\},\{t_1,\bar{t_2}\},\{\bar{t_1},t_2\},\{\bar{t_1},\bar{t_2}\}\}$. L'instance θ_1^1 du terme T_1 dans la première configuration de θ, $\theta^1 = \{t_1,t_2\}$, est $\theta_1^1 = t_1$.

Par application de la **loi jointe** selon l'approche des RBs, dans lesquels la probabilité conditionnelle d'un nœud est fonction de toutes les configurations possibles de ses noeuds parents ; et en utilisant la notation θ_i^s pour représenter une instance i d'un nœud particulier T_i telle que dans la configuration de θ^s de θ , on pourra écrire :

$$P(q,\theta^s,d,c_k) = P(d) * P(c_k) * P(q/\theta^s) * \prod_{T_i \in Q \cap (D_j \cup C_k)} P(\theta_i^s/d,c_k) \tag{6.10}$$

Par conséquent l'équation 6.9 devient :

$$P(q,d,c_k) = P(d) * P(c_k) * \sum_{\theta^s \in \theta} \left(P(q/\theta^s) * \prod_{T_i \in Q \cap (D_j \cup C_k)} P(\theta_i^s/d,c_k) \right) \tag{6.11}$$

La quantification de $P(q,d,c_k)$ revient à estimer chaque membre de la formule 6.11.

Des probabilités *a priori* sont affectées aux documents de la collection, elles sont égales à $P(d) = 1/m$, m étant le nombre total de documents de la collection. De plus, en supposant que les documents et les centres sont indépendants, l'estimation de la probabilité du dénominateur de l'équation 6.8 correspond donc à : $P(d,c_k) = P(d) * P(c_k)$

En appliquant cette dernière simplification, la probabilité $P(q/d,c_k)$ est donnée par :

$$P(q/d_j,c_k) = \sum_{\theta^s \in \theta} [P(q/\theta^s) * \prod_{T_i \in Q \cap (D_j \cup C_k)} P(\theta_i^s/d_j) * P(\theta_i^s/c_k)] \tag{6.12}$$

Ainsi la formule du calcul de l'utilité globale, en intégrant tous les centres d'intérêts, exprimés par la formule 6.6 devient :

$$EU(r_j/D) = \Psi_{k=1..n} \left[\mu_k(r_j/d,c_k) * \sum_{\theta^s \in \theta} (P(q/\theta^s) * \prod_{T_i \in Q \cap (D_j \cup c_k)} P(\theta_i^s/d_j) * P(\theta_i^s/c_k)) \right]$$

où Ψ est un opérateur d'agrégation approprié qui sera spécifié dans la section 6.7.

6.6 Estimation des distributions de probabilités

La quantification des distributions de probabilités dans le modèle, consiste à donner une signification aux arcs reliant les différents types de nœuds du réseau, et à estimer l'utilité des décisions de présentation des documents pertinents compte tenu de tous les centres d'intérêts de l'utilisateur.

Ainsi, la composante quantitative du modèle comprend des distributions de probabilités conditionnelles où pour chaque variable $x_i \in V$, est attachée une classe de probabilités

$P(X) = P(x_i/pa(x_i))$ qui est fonction de toutes les configurations possibles de ses nœuds parents $pa(x_i)$ dans le réseau G, notée θ.

L'estimation des distributions de probabilités pour quantifier chacun des membres de l'équation de l'utilité globale 6.5 est détaillée dans la section suivante.

6.6.1 Estimation de la probabilité $P(q/\theta^s)$

La probabilité $P(q/\theta^s)$ traduit que la requête Q soit satisfaite en considérant la configuration d'ordre s effectivement instanciée des termes d'indexation parents de Q.

Le calcul de $P(q/pa(q))$ est effectué sur la base d'une agrégation *Noisy-OR* proposée dans [135; 22] de la manière suivante :

$$P(q/\theta^s) = \begin{cases} 0 \; si \; (\theta^s \cap R(\theta)) = \oslash \\ \frac{1-\Pi_{T_i \in R(\theta)} nidf(T_i)}{1-\Pi_{T_i \in \theta^s} nidf(T_i)} \; sinon \end{cases} \tag{6.13}$$

Où $R(\theta)$ est l'ensemble des instances positives parmi les configurations possibles θ des termes parents du noeud Q. $nidf(T_i)$ est la formule du idf normalisée du terme T_i dans la collection.

La formule traduit le fait que plus le nombre de termes instanciés positivement dans la configuration est élevé, plus le degré de satisfaction de la requête est important. Dans le cas où tous les termes parents de la requête sont instanciés positivement alors $P(q/\theta^s) = 1$

6.6.2 Estimation de la probabilité $P(t_i/d_j, c_k)$

La probabilité $P(t_i/d_j, c_k)$ traduit la représentativité du terme T_i dans la représentation du document D_j et le centre C_k.

$$P(t_i/d_j, c_k) = P(t_i/d_j) * P(t_i/c_k) \tag{6.14}$$

La probabilité qu'un terme T_i traduit effectivement la représentativité du document et le centre d'intérêt observé peut être estimée de plusieurs manières. Nous proposons de l'estimer selon les fonctions suivantes :

$$P(t_i/d_j) = \delta + (1-\delta) * Wtd(i,j), \; \delta = 0.5 \tag{6.15}$$

$$P(t_i/c_k) = \gamma + (1-\gamma) * Wtc(i,k), \; \gamma = 0.1 \tag{6.16}$$

Avec $Wtd(i,j)$ et $Wtc(i,k)$ représentant respectivement les degrés d'importance du terme T_i dans le document D_j et le centre C_k. La formule de pondération $Wtc(i,k)$, détaillée et expérimentée, sera présentée dans le chapitre 7. On peut utiliser différents types de formules proposées dans la littérature, pour notre part, nous avons opté pour la formule $BM25$ [145]. Ainsi, $Wtd(i,j)$ est donnée par :

$$Wtd(i,j) = 0.5 * \frac{tf_{ij} \star \log(\frac{N-n_i+0.5}{n_i+0.5})}{2*(0.25+\frac{0.75*dl_j}{avg\_dl})+tf_{ij}} \tag{6.17}$$

où : n_i : le nombre de documents contenant t_i,

N : le nombre de documents dans la collection,

dl : la longueur du document d_j (nombre de termes du document),

avg_dl : la longueur moyen des documents de la collection,

tf_{ij} : la fréquence d'apparition du terme t_i dans le document d_j

6.6.3 Valeur d'utilité

Le degré d'utilité de la décision de présenter ou non un document dépend à la fois du contenu du document et des centres d'intérêts relativement à la requête en cours d'évaluation. La valeur de l'utilité élémentaire de pertinence exprime le degré de concordance entre le centre d'intérêt instancié et le document observé. Elle est liée à la décision de restituer le document D_j lorsque l'on observe le centre d'intérêt C_k. Sa valeur est d'autant plus élevée que le centre d'intérêt est similaire au document observé. La similarité document-centre peut être mesurée en se basant sur les appuis RI (vectoriel, probabiliste, etc.). Dans notre cas, on propose les deux formulations suivantes :

1. Une formule est basée sur les spécificités des relations entre les termes du document et les centres d'intérêts observés :

$$\mu_k^1 = \mu_k(r_j/d,c_k) = \frac{1+\sum_{T_i\in D_j} nidf(t_i)}{1+\sum_{T_i\in D_j - C_k} nidf(t_i)}, \in [1, 1+\sum_{T_i\in D_j} nidf(t_i)] \tag{6.18}$$

   Cette valeur sera égale à 1 lorsqu'il n'y a aucune correspondance entre le centre d'intérêt C_k et le document D_j observé. Plus il y a de termes communs entre le document et le centre d'intérêt, plus cette valeur d'utilité sera importante.

2. La seconde formule est la mesure de similarité standard du cosinus :

$$\mu_k^2 = \mu_k(r_j/d,c_k) = \frac{2*\sum_{T_i\in D_j\cap C_k} Wtd(i,j)*Wtc(i,k)}{\sqrt{\sum_{T_i\in D_j} Wtd(i,j)^2}*\sqrt{\sum_{T_i\in C_k} Wtc(i,k)^2}} \tag{6.19}$$

La valeur de l'utilité élémentaire de non pertinence associée à la décision de ne pas restituer le document D_j lorsque l'on observe le centre d'intérêt C_k, noté $\mu_k(\overline{r_j}/d, c_k)$, est donnée par :

$$\mu_k(\overline{r_j}/d, c_k) = \begin{cases} 0 \; si \; (\tau(D_j) \cap \tau(C_k) = \oslash) \\ \frac{1}{\mu(r_j/d, c_k)} \; sinon \end{cases} \tag{6.20}$$

6.7 Choix de l'opérateur d'agrégation Ψ

Le score final de pertinence d'un document observé D_j est obtenu en évaluant l'utilité globale de ce document $EU(r_j/D)$ compte tenu de la requête soumise et de l'ensemble de tous ces centres d'intérêts. Cette utilité globale, liée à la décision de restituer ou pas le document observé D_j dépend fortement des valeurs des utilités élémentaires calculées pour chaque paire document-centre d'intérêt. Elles expriment les utilités des décisions de restituer D_j en considérant un centre d'intérêt C_k de l'utilisateur à la fois.

Etant donné que l'utilisateur est représenté par plusieurs centres d'intérêts, la requête peut couvrir un ou plusieurs de ces centres. Afin de mesurer l'utilité globale du document pour l'ensemble de tous ces centres d'intérêts, on propose d'agréger les valeurs des utilités élémentaires correspondantes de manière à satisfaire les relations pouvant exister entre les centres d'intérêts de l'utilisateur selon les deux hypothèses suivantes [173] :

– ***Hypothèse 1*** : *Les centres d'intérêts de l'utilisateur sont indépendants*
  Dans ce cas, l'impact du centre d'intérêt est considéré isolément dans le calcul de l'utilité globale. En effet, le score de pertinence du document devrait être plus important pour les centres d'intérêts proches à la requête que ceux qui ne le sont pas.
  En d'autres termes, étant donné l'indépendance des centres, leur l'influence sur le processus d'inférence est fortement guidée par le centre le plus important dans le contexte de la requête. Il est donc opportun de choisir l'opérateur d'agrégation permettant de considérer l'importance mutuellement exclusive des centres, c'est-à-dire permettant de considérer l'utilité élémentaire obtenue pour le centre prédominant au détriment des valeurs d'utilité calculées pour les autres paires document-centre de moindre relation avec la recherche en cours.
  Une formulation possible de cet opérateur d'agrégation peut être donnée en maximisant les valeurs des utilités élémentaires comme suit :

$$\Psi(\mu_1, ..., \mu_n) = Max(\mu_1, ..., \mu_n) \tag{6.21}$$

– ***Hypothèse 2*** : *Les centres d'intérêts sont dépendants*
  Dans ce cas, la dépendance traduit l'existence d'un lien sémantique entre un sous ensemble de centres d'intérêts couvrant le même besoin en information que la requête. Ce lien sémantique peut s'exprimer à travers des relations de hiérarchisation des centres

d'intérêts et permet ainsi le renforcement du degré de pertinence des documents.

En effet, si l'on considère que la requête couvre un ensemble de centres d'intérêts fortement liés entre eux et que le document observé obtient une valeur d'utilité élevée pour l'un des centres d'intérêts de ce sous ensemble, alors on peut considérer ce document comme pertinent. Si, en plus les valeurs des utilités élémentaires obtenues par le document pour chacun des autres centres d'intérêts de ce sous ensemble, sont également importantes, il est judicieux de combiner ces valeurs et de tenir compte de l'impact commun de ces centres d'intérêts dans le processus décisionnel de recherche.
Ainsi, l'utilité globale du score de pertinence traduisant la décision de restituer le document sera obtenue par agrégation cumulative des valeurs des utilités élémentaires. L'opérateur d'agrégation proposé est en l'occurrence l'opérateur de la somme, définit comme suit :

$$\Psi(\mu_1 \dots \mu_n) = \sum(\mu_1 \dots \mu_n) \tag{6.22}$$

où u est le nombre total des centres d'intérêt et $\mu_k = \mu_k(r_j/d, c_k) * p(q/d, c_k)$ l'utilité élémentaire d'ordre k associée à la décision de restituer le document D_j en considérant le centre d'intérêt C_k.

6.8 Conclusion

Dans ce chapitre, nous avons présenté notre contribution à la formalisation d'un modèle d'accès personnalisé à l'information, via l'utilisation d'un diagramme d'influence, qui est une extension des réseaux Bayésiens, dédié à la résolution de problèmes liés à la prise de décision. Le diagramme d'influence permet d'avoir une représentation intuitive et expressive du processus de RI, ainsi qu'un cadre formel pour l'évaluation de pertinence d'un document vis-à-vis de la requête de l'utilisateur et de ses centres d'intérêts. Ces derniers représentent des besoins informationnels à long terme (stables) de l'utilisateur.

Ainsi, ce formalisme puissant nous a permis de modéliser les entités inhérentes à la RI classique (termes d'indexation, documents et requêtes de l'utilisateur) accompagné de ce qui confère une dimension de personnalisation au modèle : les centres d'intérêts de l'utilisateur.

Ce modèle est basé sur les diagrammes d'influence (extension de Réseaux Bayésiens). Du point de vue qualitatif, le modèle proposé traduit la structure des centres d'intérêt de l'utilisateur, des documents et index de la collection. Du point de vue quantitatif, le diagramme d'influence traduit l'influence mutuelle existante entre un besoin en information exprimé par une requête et le contexte associé, dans une situation liée à la prise de décision quant à la pertinence d'un document.

L'évaluation d'une requête y est alors vue comme un processus de propagation d'inférence ayant pour objet de maximiser l'utilité cumulée des décisions parcellaires associées aux

noeuds du diagramme. Outre le cadre formel, l'utilisation d'un DI pour modéliser le problème de l'accès personnalisé à l'information offre des atouts majeurs :

– introduire le profil de l'utilisateur comme composante, à part entière, du modèle,
– considérer l'influence du contexte sur la pertinence relative d'un document pour un utilisateur,
– ordonnancer les documents sur la base d'une fonction de pertinence cumulative tenant compte de l'utilité des décisions associées à leur présentation compte tenu des centres d'intérêts de l'utilisateur qui a émis la requête

La validation du modèle proposé sera présentée à travers les différentes expérimentations décrites dans le chapitre suivant.

7 Evaluation expérimentale du modèle d'accès personnalisé à l'information

7.1 Introduction

Dans le but de valider le modèle d'accès personnalisé présenté dans le chapitre précédent, la démarche classique en RI est d'effectuer une série d'expérimentations sur une collection de test donnée dans le cadre d'un protocole d'évaluation standard.

Le modèle d'évaluation de référence généralement adopté dans les campagnes d'évaluation en RI est le modèle de Cranfield [45]. Rappelons que ce modèle fournit une base d'évaluation comparative de l'efficacité de différents algorithmes, techniques et/ou systèmes moyennant des ressources communes : collections de test contenant des documents, des requêtes préalablement construites et des jugements de pertinence associés, des métriques d'évaluation essentiellement basées sur le rappel et la précision. C'est en l'occurrence le modèle d'évaluation adopté dans la campagne d'évaluation internationale TREC[1].

Néanmoins, ce type de campagne d'évaluation n'est pas adapté à une évaluation dans le cadre d'une recherche personnalisée. En effet, comme nous l'avons mentionné dans le chapitre 4, le manque de campagne d'évaluation standard en RI personnalisée constitue une problématique majeure dans ce domaine. L'une des principales raisons est que ce type de protocole évalue une pertinence thématique et non pas une pertinence contextuelle ou situationnelle. En effet, les tests effectués ne prennent en considération ni le contexte dans lequel se fait la recherche, ni la perception de la pertinence de l'utilisateur dans ce contexte, ni la diversité des centres d'intérêts de l'utilisateur. Ainsi, à notre connaissance il n'existe actuellement aucune évaluation de la tâche *personnalisation* dans TREC, ni de consensus commun de la communauté en RI proposant un cadre expérimental pour la personnalisation de l'accès à l'information.

Dans le but d'apporter une première solution d'une part et d'exploiter au mieux les ressources de la campagne de référence TREC d'autre part, nous proposons un cadre d'évaluation par

[1] Text Retrieval Conference, http ://trec.nist.gov

augmentation de la collection TREC par les centres d'intérêts d'utilisateurs simulés. TREC fournit des collections de test contenant un ensemble de requêtes, un ensemble de documents et des jugements de pertinence associés pour chaque requête, mais aucun élément caractérisant l'utilisateur. Notre objectif est d'exploiter ces associations de pertinence pour construire des centres d'intérêts pour des utilisateurs simulés. Dans ce sens, nous proposons d'associer des sous ensembles de requêtes traitant d'un même domaine, à un profil d'utilisateur. A partir des jugements associés, on construit automatiquement des centres d'intérêts que l'on utilise pour évaluer notre modèle d'accès personnalisé selon le protocole TREC.

La suite de ce chapitre est organisée comme suit : la section 7.2 présente notre proposition d'un protocole d'évaluation d'une tâche d'accès personnalisé à l'information selon la démarche TREC. Nous y présentons notamment la démarche adoptée pour augmenter les collections de test en y intégrant des ressources décrivant les centres d'intérêts d'utilisateur ainsi que le principe d'apprentissage-test mené pour évaluer l'efficacité de notre modèle relativement à deux référentiels (un modèle de base ne considérant pas les profils utilisateurs et un modèle d'accès personnalisé à l'information). La section 7.3 est consacrée à la présentation et discussion des résultats obtenus lors des expérimentations portant essentiellement sur l'évaluation des performances et de l'impact des différents paramètres du modèle. Un bilan issu des résultats obtenus sera donné dans la section 7.4.

7.2 Définition d'un protocole d'évaluation pour la RI personnalisée

Un protocole d'évaluation d'un système d'accès personnalisé à l'information a pour objectif d'évaluer d'une part les performances du processus d'accès intégrant le profil de l'utilisateur, et de mesurer la qualité des profils utilisateurs modélisés par le système d'autre part.

L'évaluation de la « *qualité du profil* » traduit son adéquation avec les centres d'intérêts effectifs de l'utilisateur. L'accomplissement de cet objectif nécessite idéalement des ressources humaines (vrais utilisateurs) importantes pouvant certifier que les profils appris répondent effectivement à l'expression des besoins, préférences et centres d'intérêts de l'utilisateur. Il est donc difficilement réalisable de mener une telle campagne à grande échelle. L'une des alternatives intéressantes est de pouvoir mesurer cette qualité grâce à des métriques comparatives entre les valeurs de pertinence liées aux résultats de recherche obtenus par différents profils. Le profil adéquat sera celui qui aura fourni le meilleur rendement de recherche (éventuellement en termes de rappel et de précision).

L'évaluation de « *l'impact du profil* » traduit à quel point l'intégration du profil, dans le modèle d'accès personnalisé proposé, peut apporter des améliorations dans les performances de recherche. Afin de mesurer cette évolution des performances, il est nécessaire de se baser sur un référentiel comparatif entre les modèles d'accès personnalisé dans l'absolu et de manière plus générale entre le modèle de RI personnalisée et un modèle de RI traditionnelle.

Néanmoins, peu de protocoles apportent une réponse à ces deux objectifs. En effet, la définition d'un scénario d'évaluation axé autour de l'utilisateur constitue une part importante des verrous technologiques auxquels est confronté la communauté en RI personnalisée [77].

Confronté à cette même problématique lors de l'évaluation de notre modèle d'accès personnalisé et motivé par la nécessité d'en valider l'apport, nous avons mené une réflexion sur la possibilité d'exploiter les ressources issues des campagnes d'évaluation existantes (telle que TREC) quitte à en modifier certains paramètres pour les adapter à un cadre personnalisé. Ainsi, dans le but de pallier au manque de collectif d'utilisateurs réels pour la mise en œuvre d'une campagne d'évaluation, une solution possible est de simuler leurs activités pour la construction de profils lors des expérimentations.

7.2.1 Collection de test

La collection de test de la campagne d'évaluation TREC utilisée dans ces travaux est celle des disques 1, 2 et 3. Les documents de cette collection sont issus de différents articles de presse tels que *Associate Press (AP)* and *Wall street journal (WJS)*.

La particularité de cette collection est que ses requêtes sont annotées d'un champ particulier noté « **Domain**» qui décrit un sujet d'actualité traité par la requête. Ce sont ces requêtes qui nous serviront à simuler les profils utilisateurs exploités dans le protocole d'évaluation proposé. Ci-dessous un exemple de requête extraite du domaine « *Military*» :

```
<top>
<head> Tipster Topic Description
<num> Number: 062
<dom> Domain:  Military
<title> Topic:  Military Coups D'etat
<desc> Description: Document will report a military coup d'etat,
either attempted or successful, in any country.
<smry> Summary: Document will report a military coup d'etat,
either attempted or successful, in any country.
</top>
```

La collection contient au total 12 domaines annotant les requêtes. Nous avons sélectionné aléatoirement quatre domaines d'intérêts : *Environment, Law & Government, International Relations, Military*. TREC attribut une numérotation unique à l'ensemble des requêtes pour toutes les collections qu'il fournit. Pour notre part, nous avons utilisé plus particulièrement les requêtes de la collection numérotées de 51 à 100 ($q_{51} - q_{100}$), par domaine. Chaque domaine contient le nombre de requêtes suivant : 4, 4, 5, 4, respectivement. Le tableau 7.1 donne les numéros de requêtes associées à chacun de ces domaines.

Ainsi, la collection utilisée dans notre protocole d'évaluation se caractérise par les données statistiques résumées dans le tableau 7.2.

Domaines	Requêtes associées	Numéro du domaine
Environment	59, 77, 78, 83	1
Law & Government	70, 76, 85, 87	2
International Relations	64, 67, 69, 79, 100	3
Military	62, 71, 91, 92	4

Tableau 7.1 — Numéros de requête associée aux domaines sélectionnés

Nombre de domaines	4
Nombre de documents	741670
Nombre de requêtes	17
Nombre de termes	524650
Longueur moyenne de document	193,645
Longueur moyenne de requête	3,5

Tableau 7.2 — Données statistiques de la collection de test

7.2.2 Simulation des centres d'intérêts

Cette section décrit le processus de simulation des profils utilisateurs servant à augmenter la collection de test TREC, présentée précédemment. L'idée de base est d'exploiter les associations « *Domaines - Requêtes - Documents* » comme ressources informationnelles pour l'apprentissage des centres d'intérêts des utilisateurs. En effet, tel que nous l'avons mentionné, les requêtes de la collection sont annotées d'un domaine particulier pouvant être associé à des domaines d'intérêts. De plus, TREC fournit, pour chaque requête de la collection de test, la liste des documents pertinents et non pertinents jugés par des utilisateurs réels (les assessors).

Ainsi, partant de l'hypothèse qu'à chaque domaine de la collection correspond un profil utilisateur, le processus de simulation consiste à construire pour chaque domaine contenant n requêtes, n centres d'intérêts. Ce processus de simulation se déroule de la manière suivante :

1. Pour chaque domaine k de la collection (noté Dom^k avec $k = (1..6)$), nous sélectionnons parmi les n requêtes associées à ce domaine, noté Q^k, un sous-ensemble de $n-1$ requêtes pour constituer l'ensemble d'apprentissage du centre d'intérêt, noté C_n^k.
2. A partir de cet ensemble d'apprentissage, on extrait automatiquement la liste des vecteurs[2] documents pertinents et non pertinents associés à chaque requête.
3. Partant de ces vecteurs documents, un centre d'intérêt est construit comme un vecteur de termes pondérés en appliquant l'algorithme d'apprentissage proposé dans OKAPI [145] donné par la formule $BM25$ suivante :

$$Wtc(i,k) = \log \frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R-r+0.5)} \tag{7.1}$$

[2]Le vecteur document est composé de poids associés à chacun des termes d'indexation

Où : R : le nombre de documents pertinents de la requête associé au centre C_n^k ; r : le nombre de documents pertinents contenant le terme t_i ; n : le nombre de documents contenant le terme t_i ; N : le nombre total de documents de la collection.

4. On réitère le processus n fois pour chacun des n centres d'intérêts associés au domaine k. A chaque itération, on fait varier le sous ensemble d'apprentissage pris parmi n requêtes associées au domaine k en sélectionnant une nouvelle requête à chaque fois.

Le nombre de termes constituant les centres d'intérêts ainsi obtenus est fixé à 100. Bien évidemment, lors de l'évaluation du système, les requêtes appartenant à *l'ensemble d'apprentissage* sont ignorées lors des expérimentations, ceci pour éviter que le résultat de l'évaluation ne soit biaisé par l'utilisation de requêtes ayant déjà servi à construire les centres d'intérêts. Le tableau 7.3 donne un exemple pour la construction des centres d'intérêts associés au domaine « *Environment*». Il présente les différents sous ensemble de requêtes utilisées lors de la phase d'apprentissage-test, ainsi qu'un extrait du vecteur de terme obtenu pour le premier centre. Le domaine contient 4 requêtes, on construite donc 4 centres d'intérêts.

Domaine traité	Environnement	
Requêtes associées	59 77 78 83	
Construction des centres d'intérêts		
Centre d'intérêt	Requêtes d'apprentissage	Requêtes de test
C_1^1	77 78 83	59
C_2^1	59 78 83	77
C_3^1	59 77 83	78
C_4^1	59 77 78	83
Extrait du fichier centre d'intérêt C_{11} des 4 premiers termes		
Nom-centre	Num-terme	Poids-terme
env	92	1
env	87	0.569516
env	181	0.516649
env	669	0.447064

Tableau 7.3 — Exemple de construction de centre d'intérêt

7.2.3 Stratégie de test

Dans le but de mesurer l'impact de l'intégration du profil utilisateur dans le processus d'accès à l'information nous optons pour un cadre d'évaluation TREC avec un scénario qui se base sur la méthode de la *validation croisée* [124] et ce, pour ne pas biaiser les résultats avec un seul jeu de test.

La validation croisée ou la k-fold *cross validation* est une méthode d'évaluation qui consiste à diviser la collection de test en k sous ensembles de taille égale (approximativement), d'utiliser $k-1$ sous ensembles pour l'apprentissage des centres d'intérêts dans notre cas, et le k^{ime}

sous ensemble pour le test. On réitère ensuite le processus k fois pour chacun des centres d'intérêts évalués. La figure 7.1 présente le principe de la validation croisée.

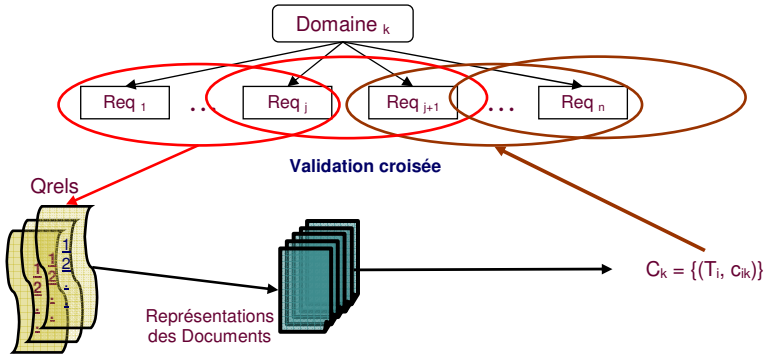


Figure 7.1 — Validation croisée pour la simulation des centres d'intérêts

Ce scénario est mis en évidence par l'algorithme suivant :

Début
Tester les requêter Q^k associées au domaine C_k
Répéter
- Subdiviser l'ensemble des n requêtes du domaine en $q_{apprentissage}$ sous-ensemble d'apprentissage de $n-1$ requêtes
et en q_{test} sous-ensemble de test contenant la n^{ime} requête
- Construction du centre d'intérêt C_k à partir de l'ensemble $q_{apprentissage}$ en en appliquant le processus de simulation.
- Tester la requête q_{test} sur notre modèle

Jusqu'à tester toutes les requêtes de Q^k
Fin

Algorithme général d'apprentissage-tests

7.2.4 Métriques d'évaluation

Les métriques d'évaluation utilisées dans ce protocole sont celles de la précision pour les X premiers documents restitués (PX) et de la MAP (Mean Average Precision) la moyenne de toute les précisions (cf. 4). Ces deux métriques permettent d'évaluer les performances de recherche selon deux aspects :

* PX : est la proportion de documents pertinents dans les X premiers documents retrouvés. Elle permet d'exprimer la satisfaction de l'utilisateur vis-à-vis des X premiers résultats pertinents. Elle constitue ainsi une mesure importante pour l'évaluation des modèles d'accès personnalisé à l'information. Dans notre cas, on retient les précisions pour les 5 et 10 premiers documents : $P5$ et $P10$.

- MAP : est la précision moyenne pour l'ensemble des documents pertinents retournés. Elle exprime la capacité du modèle à sélectionner les documents pertinents en réponse à toutes les requêtes expérimentées.

Par ailleurs, la méthode d'évaluation est faite selon le protocole TREC. Plus précisément, pour chaque requête de la collection, les 1000 premiers documents sont restitués par le système, et les valeurs de précision de $P5$, $P10$ et MAP sont calculées. On présente pour chaque requête testée les résultats pour ces mesures, puis pour chaque domaine, la moyenne des valeurs obtenues par toutes les requêtes associées. Nous comparons ensuite les résultats obtenus en utilisant notre modèle à ceux obtenus en utilisant un modèle de référence. Si le taux d'accroissement apporté par notre modèle par rapport au référentiel dépasse les 5%, on peut conclure que notre modèle apporte une amélioration significative des performances. Par contre, si le taux d'accroissement est inférieur à -5%, cela traduit une détérioration des performances de recherche de notre modèle.

7.2.5 Scénarios d'évaluation

Pour évaluer le modèle que nous proposons nous avons besoin d'un modèle de référence permettant de quantifier l'apport du profil utilisateur dans le processus d'accès à l'information. Nous pouvons comparer notre modèle à n'importe quel modèle classique de recherche d'information ne tenant pas compte des centres d'intérêts de l'utilisateur. Cependant, notre modèle étant une extension des réseaux Bayésiens, il est plus significatif de considérer comme référence les résultats obtenus avec un tel modèle. De ce fait, lors de nos expérimentations, nous avons implémenté une adaptation du modèle de recherche basé sur les réseaux Bayésiens simple (le Naïve Bayse) de [184]. Dans ce modèle la fonction de pertinence $RSV(Q,D)$, en mesurant le degré d'appariement requête-document selon la formule suivante :

$$P(Q/D_j) = \sum_{\theta^s \in \theta} [P(Q/\theta^s) * \prod_{T_i \in (Q \cap D_j)} P(\theta_i^s/D_j)] \tag{7.2}$$

Ainsi, les tests de performances sont effectués par des expérimentations comparatives entre l'évaluation de requêtes avec et sans profil utilisateur. Nous présentons dans la section suivante les résultats expérimentaux obtenus.

7.3 Expérimentations et résultats

Dans la perspective de montrer la viabilité de notre modèle et de mesurer ses performances, nous avons mené des séries d'expérimentation dont les principaux objectifs sont les suivants :

1. Evaluer l'apport de notre modèle comparativement au modèle de référence : un modèle de recherche classique basé sur les Réseaux Bayésiens (n'intégrant pas le profil utilisateur).

2. Evaluer l'impact de la fonction d'utilité pour mesurer la correspondance effective entre les documents et les centres d'intérêts.
3. Evaluer l'impact de l'opérateur d'agrégation sur les performances de recherche en comparant différents opérateurs selon les relations de dépendance entre les centres d'intérêts observés lors de l'évaluation de la requête.

Nous présentons dans les paragraphes suivants les résultats les plus significatifs obtenus.

7.3.1 Evaluation des performances du modèle

Cette section est consacrée à l'évaluation des performances de notre modèle. Sur la base de la stratégie de test définie dans notre protocole d'évaluation. Nous avons construit en premier les centres d'intérêts à partir des quatre domaines sélectionnés : « *Environment*, *International Relations*, *Law &Government*, *Military* », en appliquant le processus de simulation. A partir de ces centres d'intérêts, nous avons effectué une série de tests sur l'ensemble des requêtes associées à chaque domaine.

L'expérimentation consiste à exécuter ces requêtes sur deux modèles de recherche : notre modèle et le modèle Bayésien Simple (Naïve Bayse, NB). L'objectif est de démonter que notre modèle d'accès personnalisé basé sur les Diagrammes d'influence[3] est plus performant que le modèle de recherche classique basé sur les RBs. Cette expérimentation va permettre d'agréer notre proposition de considérer le processus d'accès à l'information comme un processus décisionnel. La configuration des paramètres de notre modèle utilisés lors de cette série d'expérimentation est tel que nous avons utilisé la somme $\sum$ comme opérateur d'agrégation et la fonction d'utilité donnée par la formule suivante :

$$\mu_k^1 = \mu_k(r_j/d, c_k) = \frac{1+\sum_{T_i \in D_j} nidf(t_i)}{1+\sum_{T_i \in D_j - C_k} nidf(t_i)}, \in [1, 1+\sum_{T_i \in D_j} nidf(t_i)]$$

Le tableau 7.4 présente les résultats expérimentaux obtenus pour chacun des domaines testés par ces deux modèles de la manière suivante :

1. Pour chaque requête de chacun des domaines testés, nous donnons les valeurs aux points de précision $P5$, $P10$ et la valeur de la MAP.
2. Pour chaque domaine, on calcule la moyenne sur toutes ses requêtes les valeurs obtenues en 1. pour les mêmes métriques ($P5$, $P10$ et MAP).
3. Un taux d'accroissement est calculé sur la base des moyennes obtenues en 2. pour chaque domaine. Ce taux d'accroissement est obtenu de manière générale pour deux variables A et B mesurant le pourcentage de C par

$$\%C = \frac{B-A}{B} \times 100$$

[3]Rappelons que les DI sont une extension des Réseaux Bayésiens dédié aux problèmes décisionnels

Dans notre cas, le variable *A* représente les valeurs moyennes obtenues par le modèle de base, la variable *B* représente ceux de notre modèle et la variable *C* le pourcentage d'amélioration de notre modèle par rapport au modèle de base.

	Modèle de base NB			**Notre modèle**		
Environment	*P5*	*P10*	*MAP*	*P5*	*P10*	*MAP*
59	0,40	0,40	0,01	**0,80**	**0,80**	**0,05**
77	0,80	0,70	**0,39**	**1,00**	**1,00**	0,25
78	1,00	1,00	**0,75**	**1,00**	**1,00**	0,35
80	0,00	0,10	**0,03**	**0,40**	**0,20**	0,01
Moyenne	**0,55**	**0,55**	**0,29**	**0,80**	**0,75**	**0,17**
Taux Accroissement				**+45%**	**+36%**	**-41%**
International Relations	*P5*	*P10*	*MAP*	*P5*	*P10*	*MAP*
64	0,20	0,20	0,18	**0,80**	**0,60**	**0,24**
67	0,00	0,10	0,00	**0,40**	**0,30**	**0,01**
69	0,20	0,20	0,08	**1,00**	**1,00**	**0,47**
79	0,00	0,00	0,00	**1,00**	**0,60**	**0,08**
Moyenne	**0,10**	**0,12**	**0,06**	**0,80**	**0,62**	**0,20**
Taux Accroissement				**-37%**	**-7%**	**-1%**
Law &Government	*P5*	*P10*	*MAP*	*P5*	*P10*	*MAP*
70	0,60	0,60	0,42	**1,00**	**1,00**	**0,65**
76	**0,60**	**0,70**	0,08	**0,60**	0,30	**0,09**
85	**0,60**	**0,80**	**0,21**	**0,60**	0,70	0,16
87	0,20	0,20	0,00	**1,00**	**0,60**	**0,05**
Moyenne	**0,50**	**0,57**	**0,17**	**0,80**	**0,65**	**0,24**
Taux Accroissement				**+60%**	**+13%**	**+35%**
Military	*P5*	*P10*	*MAP*	*P5*	*P10*	*MAP*
62	0,20	0,40	0,33	**0,80**	**0,80**	**0,80**
71	**1,00**	**1,00**	**0,80**	0,20	0,20	0,20
91	0,00	0,00	0,00	**0,80**	**0,60**	**0,60**
92	0,00	0,00	0,00	**0,80**	**0,60**	**0,60**
Moyenne	**0,30**	**0,35**	**0,28**	**0,50**	**0,42**	**0,42**
Taux Accroissement				**+66%**	**+21%**	**+50%**

Tableau 7.4 — Résultats expérimentaux par domaine

La première constatation au vu de ces résultats est que notre modèle d'accès personnalisé à l'information est efficace et réalise des améliorations de performances significatives comparativement au modèle de base aux points de précision $P5$ et $P10$ pour trois domaines sur les quatre testés. On note également une amélioration significative de la *MAP* pour les deux domaines (*Law& Government, Military*), mais diminue aussi significativement pour deux

autres domaines : (*Environment, International Relations*). Néanmoins, ces résultats sont acceptables compte tenu des valeurs de précision à $P5$ et $P10$, valeur de la proportion des premiers documents pertinents restitués en haut de liste par notre modèle.

Lorsqu'on analyse les taux d'accroissement pour chaque requête, on remarque que ce degré d'amélioration change d'une requête à une autre. Ceci dépend probablement, d'une part, des corrélations entre les centres d'intérêts de l'utilisateur simulés et le domaine de la requête (exprimés dans notre modèle par la fonction d'utilité) et d'autre part du niveau de performance du modèle de base.

7.3.2 Evaluer l'impact de la fonction d'utilité

L'objectif de cette section est de mesurer l'efficacité des deux fonctions d'utilité proposées dans notre modèle. La fonction d'utilité traduit le degré de correspondance entre un document et un centre d'intérêt utilisateur. Sa capacité à exprimer effectivement ce type de relation est d'autant plus importante qu'elle constitue un facteur discriminant lors du processus d'évaluation de la pertinence du document. Rappelons ci-dessous les deux formules proposées :

1. Une formule est basée sur les spécificités des relations entres les termes du document et le centre d'intérêt observé :

$$\mu_k^1 = \mu_k(r_j/d, c_k) = \frac{1+\sum_{T_i \in D_j} nidf(t_i)}{1+\sum_{T_i \in D_j - C_k} nidf(t_i)}, \in [1, 1+\sum_{T_i \in D_j} nidf(t_i)]$$

2. La seconde formule expérimentée est la mesure de similarité standard du cosinus :

$$\mu_k^2 = \mu_k(r_j/d, c_k) = \frac{\sum_{T_i \in D_j \cap C_k} Wtd(i,j) * Wtc(i,k)}{\sqrt{\sum_{T_i \in D_j} Wtd(i,j)^2} * \sqrt{\sum_{T_i \in C_k} Wtc(i,k)^2}}$$

Les résultats expérimentaux obtenus pour les deux utilités sont présentés dans le tableau 7.5. Les résultats du modèle de base correspondent à ceux données par le modèle Bayésien simple. Ce tableau donne pour chaque domaine la moyenne des précisions obtenues sur l'ensemble des requêtes associées aux points de précision $P5$ et $P10$. Les résultats sont présentés uniquement pour ces mesures dans le but de déterminer la capacité de la fonction d'utilité à spécifier les relations *Document-Centre d'intérêt* pour les premiers documents pertinents retournés.

La première remarque au vu de ces résultats, est que notre modèle apporte une amélioration significative des valeurs de précision comparativement au modèle de base, et ce quelque soit la fonction d'utilité utilisée pour les deux mesures dans la majorité des domaines. On constate également que la fonction de l'utilité μ_k^1 a un impact réel sur les performances du

	Utilité μ_k^1		Utilité μ_k^2		Modèle de base	
Domains	P5	P10	P5	P10	P5	P10
Environment	**0,6000**	**0,5750**	0,5500	0,5143	0,3500	0,2750
Law & Government	**0,5500**	**0,5000**	0,4357	0,3750	0,3000	0,3500
International Relations	0,2400	**0,2600**	**0,2501**	0,2324	0,2400	0,3200
Military	**0,4000**	**0,3250**	0,3617	0,3104	0,2000	0,1500

Tableau 7.5 — Résultats expérimentaux de l'impact de la fonction d'utilité

modèle par rapport à la fonction μ_k^2. Plus particulièrement, on note un taux d'accroissement global pour tous les domaines de $47,34\%$ à $P5$ pour la fonction μ_k^1 et de $41,96\%$ à $P5$ pour la formule de l'utilité μ_k^2 par rapport au modèle de base.

Si l'on analyse finement ces résultats par domaine, on observe un taux d'accroissement pour les domaines « *Environment* » et « *Law & Government* » de $71,43\%$ et de $83,33\%$ à $P5$ en utilisant la fonction μ_k^1, respectivement. Dans le cas de l'utilité de la fonction μ_k^2, les taux d'accroissement pour ces domaines sont, relativement plus faibles, égaux à $57,14\%$ et à $45,24\%$ à $P5$, respectivement.

Ces résultats suggèrent l'importance de la fonction d'utilité μ_k^1. Ceci confirme notre intuition que l'utilité ne peut être exprimée par une simple mesure de similarité. Elle doit exprimer le lien de dépendance entre un document et le centre d'intérêt, en tenant compte de l'importance des termes représentant le document, le centre d'intérêt et également la requête. L'importance de ces termes est donnée par les valeurs de pondération reportées dans la fonction d'utilité. En effet, dans le cas où un document ne correspond à aucun centre d'intérêt de l'utilisateur[4], il est important de pouvoir traduire cette information, à travers la fonction d'utilité, sans pour autant pénaliser la valeur globale de pertinence attribuée à ce document. Il est tout à fait probable que le document soit pertinent vis-à-vis de la requête même s'il ne correspond à aucun centre d'intérêt. Dans ce cas, si l'utilité est représentée par une simple mesure de similarité, tel que le cosinus, sa valeur sera nulle. De ce fait, la valeur de pertinence *RSV* tendra vers zéro,

Nous retenons pour les expérimentations suivantes la fonction d'utilité μ_k^1.

7.3.3 Comparaison des opérateurs d'agrégation

Le score final de pertinence d'un document est mesuré par agrégation de toutes les valeurs des utilités élémentaires calculées pour ce document et pour chacun des centres d'intérêts de l'utilisateur observé, lors du processus d'évaluation de la requête. Nous avons proposé deux opérateurs d'agrégation sur la base d'hypothèse de dépendance et indépendance des centres d'intérêts. Ces opérateurs sont l'opérateur de *somme* dans le cas où les centres sont dépendants et l'opérateur du *maximum* dans le cas d'indépendance [173].

[4]Cas d'un nouveau besoin en information de l'utilisateur jamais rencontré lors de ses recherches

Les domaines de la collection sont supposés traiter d'un domaine d'intérêt lié à un sujet d'actualité abordé par les articles de presse dont est issue la collection. En explorant le contenu des documents et les descriptions des requêtes associées aux domaines, nous avons sélectionné quatre domaines d'intérêts de la collection à savoir «*Environment, Law & Government, International Relations et Military* ». Ainsi, les relations de dépendance et indépendance entre ces domaines sont supposés comme suit :

– Les domaines «*International relations* » et «*Law&Government* » sont dépendants.
– Les domaines «*Environment* » et «*Military* » sont indépendants.

Les expérimentations menées pour évaluer l'impact des opérateurs d'agrégation proposés ont été réalisées en testant les deux opérateurs sur les quatre, pris deux à deux en considérant les hypothèses posées ci-dessus.

Les résultats obtenus sont présentés dans le tableau 7.6 pour le cas de domaines dépendants, et dans le tableau 7.7 pour le cas de domaines indépendants. Pour chaque domaine, on présente la moyenne des résultats obtenus sur l'ensemble des requêtes associées.

	Opérateur $\sum$			**Opérateur** *Max*		
Domaines	P5	P10	MAP	P5	P10	MAP
Environment	0,30	0,25	0,06	**0,80**	**0,75**	**0,17**
Military	0,28	0,40	0,04	**0,43**	**0,50**	**0,10**

Tableau 7.6 — Résultats d'agrégation de domaines indépendants

	Opérateur $\sum$			**Opérateur** *Max*		
Domaines	P5	P10	MAP	P5	P10	MAP
International Relations	0,50	0,55	0,18	**0,80**	**0,62**	**0,20**
Law & Government	0,60	0,50	0,18	**0,80**	**0,65**	**0,23**

Tableau 7.7 — Résultats d'agrégation de domaines dépendants

Les résultats expérimentaux indiquent pour les deux types de relations entre domaines que l'opérateur du *maximum* surpasse l'opérateur de la *somme* aux points de précision $P5$, $P10$ et la *MAP*.

Ces résultats contrastent nos hypothèses de départ sur les relations entre les domaines. Rappelons que l'opérateur du *maximum*, appliqué sur un ensemble d'attributs traduit une disjonction de valeurs, et que l'opérateur de la *somme*, une conjonction. Néanmoins, on remarque que ces caractéristiques ne sont pas reportées dans les résultats expérimentaux. Ce qui implique finalement que les quatre domaines sélectionnés sont indépendants.

Ceci peut être expliqué d'une part par le fait que les relations telles qu'elles sont définies peuvent certes traduire les liens entres les domaines, mais pas forcement les liens entre les centres d'intérêts construits à partir des documents associés aux requêtes annotées par le domaine. D'autre part, les suppositions faites sur les relations entre les domaines, ne peuvent être uniquement basées sur la thématique abordée.

Une première solution consiste à définir des mesures de corrélation entre les documents associés aux requêtes des domaines. En outre, on peut également définir ces relations, non vis-à-vis du domaine, mais directement entre les centres d'intérêts construits et qui sont par ailleurs sélectionnés lors de l'évaluation de la requête. Ainsi, on pourra appliquer l'opérateur adéquat selon les relations établies sur la base des corrélations entre centres d'intérêts candidats à l'évaluation de la requête.

7.4 Conclusion

Nous avons présenté dans ce chapitre le dernier volet de notre contribution qui consiste en la définition d'un cadre d'évaluation approprié pour l'accès personnalisé à l'information et l'évaluation de notre modèle. Nous avons alors appliqué ce cadre pour évaluer précisément notre modèle et avons présenté les résultats expérimentaux obtenus. Le cadre proposé a l'intérêt de réutiliser les ressources de la campagne d'évaluation standard TREC.

D'une manière générale, les expérimentations effectuées pour mesurer la viabilité de notre modèle décrit dans ce chapitre peuvent être répertoriées en deux classes :

1. la première classe consiste à mesurer l'efficacité du modèle proposé à intégrer le profil de l'utilisateur dans le processus d'accès à l'information. Nous comparons dans ce type d'expérimentations les performances du modèle d'accès personnalisé à celles d'un modèle classique de RI. La précision moyenne que nous obtenons est supérieure à celle obtenue par modèle Bayésien simple de plus de $40,55\%$. Ces résultats nous permettent de conclure que notre modèle est efficace et performant pour personnaliser l'accès à l'information.
2. la seconde classe s'articule autour de l'impact des différents paramètres de notre modèle sur les performances de recherche. Nous avons mesuré l'apport et la validation de la fonction d'utilité appropriée pour mesurer la corrélation entre document et centre d'intérêt. Ainsi, la fonction d'utilité qui contribue à de meilleures performances de notre modèle ne pénalise pas la pertinence des documents lorsqu'il n'y a pas de corrélation avec le profil de l'utilisateur.

De plus, nous avons évalué l'impact de deux opérateurs d'agrégation. Un opérateur de *somme* traduisant la dépendance entre les centres d'intérêts et un opérateur de *maximum* dans le cas d'indépendance de centres. Les résultats obtenus sur la collection sont cependant à l'encontre de nos hypothèses de départ. Ce qui ouvre des perspectives intéressantes à suivre pour exploiter des mesures de corrélation entre les domaines d'intérêts associés aux requêtes de la collection. Finalement, nous renforçons l'idée de l'importance des domaines d'intérêts annotant les requêtes de cette collection de tests.

Conclusion générale

Les travaux présentés dans ce manuscrit se situent dans le contexte général des systèmes d'accès à l'information et plus particulièrement dans le cadre de la personnalisation de l'accès à l'information.

Nos contributions, à travers cette thèse investissent les trois questions critiques posées par un processus d'accès personnalisé à l'information : **modélisation de l'utilisateur**, **modélisation de l'accès** et **évaluation du modèle d'accès**.

1. Dans notre approche, la modélisation du profil utilisateur s'effectue à travers deux dimensions informationnelles représentants ses " *historiques de recherches* " et ses " *centres d'intérêts* ". La dimension *historique* est vue comme une source d'information, évoluant lors des différentes sessions de recherche de l'utilisateur, à partir desquelles on fait émerger les centres d'intérêts de l'utilisateur. Plus précisément, on détermine des périodes d'apprentissage qui définissent des jalons pour l'extraction de centres d'intérêts à court terme, qualifiés de *contextes d'usage*, à partir des informations agrégées dans l'historique de recherche. L'évolution de la dimension " *Centres d'intérêts* " est alors basée sur une mesure de corrélation des rangs qui évalue le degré de changement entre contextes d'usage associés à des périodes successives. En outre, plutôt que d'exploiter l'importance intrinsèque des termes dans les documents, on propose de décliner l'importance relative des termes selon le profil de l'utilisateur lors de l'agrégation des informations collectées à partir de ses sessions de recherche successives.
   La collecte de ces informations lors des sessions de recherche successives est réalisée en appliquant d'une approche d'acquisition implicite des documents visités par l'utilisateur. Ces documents sont collectés par estimation d'un degré d'intérêt implicitement inféré du comportement de l'utilisateur lors de sa session de recherche. Notre travail a abouti au développement d'un outil $\mathcal{W}eb\ Cap$ implémentant l'algorithme d'inférence implicite proposé. L'évaluation expérimentale des performances de notre approche, nous a permis de constater une capacité égale à 80% d'acquisition des documents pertinents de manière implicite comparativement à l'approche explicite.
2. Le second volet de notre contribution porte sur la caractérisation de l'accès personnalisé à l'information comme un problème décisionnel à travers un modèle inférentiel basé sur une extension des réseaux Bayésiens en l'occurrence les diagrammes d'influence. Le modèle proposé permet de formaliser le raisonnement lié à la prise de décision quant

à la pertinence des documents compte tenu du profil de l'utilisateur d'une part et de la requête d'autre part. Le modèle est représenté par deux composantes pour codifier cette connaissance : qualitative et quantitative. La composante qualitative du modèle traduit la structure des centres d'intérêts de l'utilisateur, des documents et index de la collection. La composante quantitative, traduit la mutuelle influence existante entre un besoin en information exprimé par une requête et un contexte associé, dans une situation liée à la prise de décision quant à la pertinence d'un document. La valeur de pertinence d'un document est exprimée à l'aide de l'utilité de la décision liée à sa présentation. L'évaluation d'une requête y est alors vue comme un processus de propagation d'inférence ayant pour objet de maximiser l'utilité cumulée des décisions parcellaires associées aux noeuds du diagramme.

3. Le troisième volet de notre contribution consiste en la définition d'un cadre d'évaluation approprié pour l'accès personnalisé à l'information. Nous avons alors appliqué ce cadre pour évaluer précisément notre modèle et avons présenté les résultats expérimentaux obtenus. Le cadre proposé a l'intérêt de réutiliser les ressources de la campagne d'évaluation standard TREC. Notre objectif est d'exploiter les associations de pertinence (document-requête), existantes dans les collections TREC, pour construire des centres d'intérêts des utilisateurs simulés. Plus précisément, nous simulerons les profils des utilisateurs à partir de requêtes issues des disques 1, 2 et 3 de TREC. Le choix de cette collection a été motivé par le fait que ces requêtes sont décrites par un champ particulier qui spécifie leurs domaines respectifs et qui nous permet de simuler les centres d'intérêts de l'utilisateur. Dans ce sens, nous proposons d'associer des sous ensembles de requêtes traitant d'un même domaine, à un profil d'utilisateur. A partir des jugements associés, on construit automatiquement des centres d'intérêts que l'on utilise pour évaluer notre modèle d'accès personnalisé selon le protocole TREC. Le procédé de construction des centres d'intérêts est basé sur le principe de la validation croisée. Les expérimentations menées portant essentiellement sur l'évaluation des performances et de l'impact des différents paramètres du modèle. Les résultats obtenus montrent que notre modèle est efficace et performant pour personnaliser l'accès à l'information. Il apporte une amélioration des performances de recherche de $40,55\%$ comparativement au modèle d'accès à l'information classique en l'occurrence le modèle des réseaux Bayésiens simple.

Les résultats obtenus sont encourageants et ouvrent des perspectives intéressantes, que nous présentons dans ce qui suit.

Perspectives

De nombreuses perspectives découlent de chacune nos propositions. Elles portent essentiellement sur plusieurs points :

1. **Perspectives consternants la modélisation du profil utilisateur**
   - Le plus important porte sur la définition des périodes d'évolution des centres d'intérêts. En effet, la variation des centres d'intérêts de l'utilisateur, décelée à travers les requêtes qu'il a émis, ne présente pas forcément des régularités prévisibles ; ainsi, la méthode statistique proposée serait confrontée à un risque d'erreur difficilement mesurable. Même si ce risque pourrait être amoindri en réduisant au mieux ces périodes, une perspective intéressante est de mener une réflexion plus poussée sur un compromis entre les différents paramètres qui régulent l'évolution des centres d'intérêt d'un utilisateur.
   - A court terme, on envisage d'étendre les fonctionnalités de l'outil $\mathcal{W}eb\,Cap$ en intégrant des modules de construction et d'inférence des centres d'intérêts de l'utilisateur suivant le procédé de modélisation proposé.
   - Notre perspective à long terme consiste à étendre ses fonctionnalités afin de pouvoir intégrer le modèle d'accès personnalisé proposé.
2. **Perspectives concernant le modèle d'accès personnalisé décisionnel**
   - l'un de nos objectifs consiste à explorer de nouvelles mesures de corrélation entre les documents associés aux requêtes des domaines.
   - En outre, nous envisageons d'explorer d'autres opérateurs d'agrégation basés sur des nouvelles mesures définies, non vis-à-vis du domaine, mais directement entre les centres d'intérêts construits et qui sont par ailleurs sélectionnés lors de l'évaluation de la requête.
3. **Perspectives concernant le protocole d'évaluation**
   - A court terme, notre objectif est d'étendre le processus d'apprentissage à plusieurs centres d'intérêts pour un même utilisateur et par conséquent à un nombre plus élevé de requêtes, puis d'en évaluer l'impact sur la taille des données d'apprentissage d'une part, et des performances de recherche d'autre part. Dans la même perspective, nous envisageons d'étendre la taille des requêtes pour mesurer leur impact sur les performances de notre modèle.
   - A long terme, notre objectif est de paramétrer note modèle, sur la base de l'ensemble des résultats obtenus des perspectives envisagées, pour qu'il apporte des améliorations comparativement à un modèle d'accès personnalisé à l'information. En l'occurrence un modèle basé sur le réordonnancement des documents proposé par [166].

A Typologie des indicateurs d'intérêts pour l'observation du comportement

Le comportement observable de l'utilisateur est défini par les actions qu'a un utilisateur face aux résultats fournis par le système d'accès à l'information. Il faut cependant distinguer entre les comportements observables de l'utilisateur et les techniques qui permettent de les observer que l'on nomme communément *les indicateurs implicites d'intérêts*. Plusieurs indicateurs ont été recensés dans la littérature, tels que : la durée de lecture, le nombre de clicks de la souris, les défilements, le marquage, l'impression, la sauvegarde, la sélection, etc.

Cette annexe présente un état de l'art portant sur les indicateurs les plus utilisés. Nous donnons en premier lieu un aperçu des premiers travaux de référence. Nous présentons par la suite chacun des comportements retenus dans notre contribution ainsi que les indicateurs associés. Une synthèse de tous les travaux sera présentée à la fin de cette annexe dans un tableau récapitulatif.

A.1 Typologies des comportements

Les premiers travaux de recherche basés sur l'acquisition implicite de données utilisateur ont été menés dans le cadre du système *InfoScope* [169] : un système de filtrage d'information de groupe de discussion sur le *web* (USENET). Ce système combine les deux techniques d'acquisition pour construire le profil de l'utilisateur. Il identifie trois sources de feedback implicite : lorsque l'utilisateur lit ou ignore un message, lorsqu'il sauvegarde ou efface un message ou lorsqu'il le transmet ou pas. En 1994, [129] ont mené des expérimentations sur l'efficacité d'un système de filtrage d'information en se basant sur l'observation du temps de lecture. Ils ont constaté qu'il y a une forte corrélation entre le temps de lecture et le feedback explicite. Reprenant les travaux de [129], [101] ont proposé d'autres observations du comportement incluant l'impression et l'envoi de mail comme sources additionnelles pour le feedback implicite.

Dans ce contexte de recherche, [44] et [88] ont tenté de définir des indicateurs d'intérêts implicites. Pour cela, ils ont modifié et développé un navigateur qui détecte les différentes activités de la souris, du clavier et des barres de défilement manipulées par l'utilisateur lors de sa recherche. Néanmoins, ces deux travaux aboutissent à une contradiction. Claypool stipule que les clics de souris n'apportent pas de plus value aux estimations implicites, par contre, l'activité sur les barres de défilement à plus long terme a plus de signification. De son côté Jung affirme le contraire, que le nombre de *clics* de la souris est un bon indicateur implicite et que les barres de défilement ne le sont pas. Ceci est explicable, par le fait que les expérimentations et les tests ne s'inscrivent pas dans un cadre d'évaluation commun. Ceci relève d'une problématique majeure de l'évaluation expérimentale de la RI personnalisée. Confronté à cette incohérence de résultats, [93] ont testé d'autres indicateurs pour tenter d'apporter une clarification, telles que la distance dans le mouvement de la souris et la distance dans le mouvement des barres de navigation.

En se basant sur ces travaux, nous présentons dans ce qui suit, les comportements observables de l'utilisateur, jugés les plus importants par la communauté de RI ainsi que les indicateurs associés.

A.1.1 Comportement de « *Lecture* »

La lecture est un comportement trivial, considérée comme premier observable lors de l'activité de recherche de l'utilisateur. En effet, si un document a été lu, cette action caractérise un intérêt probable de l'utilisateur pour ce document [169]. Pour observer ce comportement il existe plusieurs indicateurs : la durée de lecture, l'activité de la souris, la barre de défilement (scrollbar), la sélection de texte.

A.1.1.1 La durée de lecture

La durée de lecture est considérée comme le plus important des indicateurs du comportement de lecture. Plusieurs travaux ont porté sur l'étude de la corrélation entre cet indicateur et le feedback explicite, pour mesurer son efficacité à traduire le jugement de pertinence de l'utilisateur. Pour cela ils ont développé des systèmes qui permettent d'enregistrer la durée de lecture des documents ainsi que les jugements explicites des utilisateurs concernant chaque document, puis ils procèdent à l'analyse des résultats avec des méthodes statistiques (analyse de la variance, régression linéaire et non linéaire, etc.).

Dans ce contexte, [129] ont dirigé une étude sur six semaines avec 8 utilisateurs pour déterminer si la préférence des messages USENET se reflétait dans la durée de lecture des messages. Les résultats ont montré une corrélation positive entre la durée de lecture et le feedback explicite exprimés par ces utilisateurs. En effet, les résultats de rappels et précision obtenus lors d'une simulation de filtrage d'information par des profils construits à partir de messages jugés implicitement pertinents par l'utilisateur (pertinence relative à une durée de lecture supérieure à 20 secondes) sont meilleurs que ceux obtenus par une évaluation ex-

plicite de pertinence. Des conclusions similaires ont également été obtenues par [101; 44]. D'autres expérimentations menées par [192] ont démontré que la durée de lecture était un indicateur discriminant entre le jugement implicite de pertinence et de non pertinence associés aux documents. En effet, les utilisateurs passent plus de temps à lire les documents qu'ils jugent pertinents que les non pertinents [94].

Cependant, lorsque [90] ont tenté de reproduire les résultats de Morita et Shinoda [129] ils ont trouvé qu'il n'y avait pas de différence entre la durée de lecture des documents jugés pertinents et ceux jugés non pertinents. Ces résultats suggèrent que la durée de lecture peut être un indicateur efficace dans le cadre de certaines tâches, mais pas pour toutes. En effet, on trouve dans [92] une étude sur la corrélation entre la tâche de recherche et la durée de lecture. Le but de cette étude est de déterminer pour quelles tâches : soumission de requête, navigation simple ou complexe, l'indicateur « *durée* » serait le plus efficace. Ces résultats sont conformes à ceux de Kelly & Berklin [90]. Ces conclusions peuvent être expliquées par le fait que, pour juger de la pertinence d'un document, il faut le lire en entier. De ce fait, la durée est la même pour les documents pertinents et non pertinents. Les résultats concernant les deux dernières tâches combinées, suggèrent que lorsque l'utilisateur cherche une part d'information, il passe plus de temps sur les documents pertinents. Ils ont aussi montré que plus la tâche est complexe, plus l'indicateur « durée » est efficace.

A.1.1.2 L'activité de la souris

Elle tend à déterminer l'activité de lecture réelle de l'utilisateur. De nombreuses études ont été menées dans le but de déterminer la corrélation entre l'activité de la souris et l'observation du comportement lecture. [69] ont développé un outil permettant de collecter des informations concernant l'activité de la souris durant une période de temps appelé « temps actif », qui a été fixée à 20mn. Dans cette expérience, ils ont considéré cette activité comme étant le passage de la souris sur un lien dans la page, ou bien la navigation dans un menu. Ils ont essayé de prédire l'activité future de la souris en ayant des informations sur son activité actuelle. Bien qu'ils n'aient pas pu trouver une forte corrélation statistique entre l'activité future et actuelle de la souris, ils ont conclu que l'erreur pouvait bien prévenir de la difficulté à détecter l'activité de la souris lors de cette expérience. L'activité de la souris comporte donc deux principaux indicateurs implicites d'intérêts : les mouvements de la souris et le nombre de ses clics, décrits ci-dessous.

– Le mouvement de la souris permet de détecter la présence d'une activité sur une page donnée. Elle se mesure par la distance parcourue de la souris sur l'écran de l'utilisateur. La mesure de distance proposée par [93] suppose que plus cette distance est grande, plus l'intérêt de l'utilisateur pour la page lue est grand. Les résultats expérimentaux ont montré l'efficacité de prédiction de cet indicateur pour l'intérêt de pertinence de l'utilisateur, au même degré que l'indicateur « durée de lecture ».

Cette distance est calculée par la formule suivante :

$$dist\_mouv\_souris(pixels) = \sum_{i=0}^{t-1} Dist(P(t_i) - P(t_i - t_{i-1})) \quad (A.1)$$

Où : *Dist* est la distance euclidienne, t représente la durée d'activation de la fenêtre, $p(t_i)$ la position de la souris par rapport à l'axe des X et des Y. L'intervalle de temps $t_i - t_{i-1}$ est de 100 ms.
Le système *KixBrowser* [88] utilise également une distance semblable pour mesurer les mouvements de la souris, calculée selon la formule suivante :

$$dist\_mouv\_souris = |souris_x - old_x| + |souris_y - old_y| + dist\_mouv\_souris \quad (A.2)$$

Où : old_x et old_y contiennent les coordonnées précédentes de la souris, par rapport à l'axe des X et des Y, afin de calculer la distance. Aussi loin que l'utilisateur bouge la souris dans la fenêtre active, la distance est accumulée.

– Le nombre de clics de la souris permet de traduire l'activité de navigation de l'utilisateur à travers les liens hypertextes. Cet indicateur a été jugé performant dans le cadre du système *KisBrowser* [88]. [93] ont démontré que son efficacité était aussi performante que celle de la durée et des mouvements de la souris. Cependant, les résultats expérimentaux menés dans le cadre du système *Curious Browser* [108] n'ont pas abouti aux mêmes conclusions.

A.1.1.3 La barre de défilement (scrollbar)

Elle est souvent utilisée pour faire défiler le contenu des pages dépassant les dimensions du moniteur de l'utilisateur. Lorsque l'utilisateur juge qu'un document est intéressant, il ne peut le lire entièrement qu'en visualisant son contenu. Cette activité peut être observée grâce à deux indicateurs : Le mouvement du scrollbar en utilisant la souris, et le mouvement du scrollbar en utilisant les différentes touches de direction du clavier $(\uparrow, \downarrow, \leftarrow, \rightarrow)$. Lors du développent de *Curious Browser*, Le et Waseda [108] ont obtenu de bonnes performances de recherche en combinant ces indicateurs par rapport aux jugements explicites de l'utilisateur.
.

A.1.1.4 La sélection de texte

Elle s'effectue lors de la lecture d'un document. L'utilisateur sélectionne souvent des parties du document dans le but de les copier ou simplement pour mieux les visualiser. Cette activité traduit l'intérêt de l'utilisateur pour ce document. Basé sur cette hypothèse, plusieurs travaux ont entamé une série de tests qui ont donné différents résultats.

Jung [88] a analysé cet indicateur sur la base du calcul de la taille du texte sélectionné en fonction de la « distance Euclidienne » entre deux points (le point du début de la sélection et celui de la fin). Cependant, l'inconvénient de cette mesure est que le texte sélectionné horizontalement ne sera pas pris en compte lorsque la souris se déplace verticalement. Pour y remédier, [93] ont proposé une nouvelle méthode de calcul. Ils supposent qu'un caractère est représenté par 5 pixels, et que chaque ligne contient 80 caractères. La distance entre deux lignes est en moyenne de 20 pixels. La formule proposée est la suivante pour calculer de la taille du texte sélectionné :

$$Taille\_Text = \sum_{j}^{E}(\frac{DistY_j}{20*80} + \frac{DistX_j}{5}) \tag{A.3}$$

Avec, E : le nombre de sélections, $DistY$ et $DistX$ la distance verticale et horizontale respectivement entre deux points.

A.1.2 Comportement de « *Sauvegarde* » , « *Impression* » et « *Annotation* »

Différents autres comportements peuvent être observés lors des sessions de recherche des utilisateurs, pour inférer leur degré d'intérêt pour un document pertinent. En effet, un document peut être sélectionné en le sauvegardant, en l'imprimant, en l'annotant ou en l'envoyant par *Email*. Ces évènements traduisent l'intérêt suscité de l'utilisateur pour ce document. Cependant, l'observation de ces comportements doit absolument être combinée avec d'autres comportements et indicateurs pour être exploitable (par exemple : la lecture et l'impression, la durée de lecture et les défilements). En effet, un utilisateur peut sauvegarder ou imprimer un document sans l'avoir lu, par faute de temps, et se rendre compte ultérieurement qu'il n'est pas pertinent en réponse à ses besoins en informations.

[37] combine la durée de lecture avec le nombre de visites d'une page pour estimer le temps total où l'utilisateur consulte une page. D'autres travaux se focalisent sur les annotations effectuées par l'utilisateur [195]. Les utilisateurs annotent souvent les pages *web* qu'ils trouvent intéressantes, pour faciliter leur consultation ultérieure. Dans le cadre du système *Letizia*, [110] utilise différents niveaux d'annotation pour suggérer différents degrés d'intérêts. Des expérimentations qu'il a effectuées, il en déduit que : (1) la sauvegarde d'une référence d'un document implique un grand intérêt de l'utilisateur,

(2) suivre un lien implique un intérêt provisoire,

(3) revisiter une page *web* indique un intérêt croissant, (4) et que passer sur un lien indique qu'il n'y a pas d'intérêt pour ce document à moins qu'il soit sélectionné plus tard.

A.2 Synthèse

Nous présentons dans le tableau A.1 une synthèse des différents comportements observables ainsi que les indicateurs associés. Nous donnons pour chacun les conclusions sur leurs performances obtenues par les expérimentations effectuées dans la littérature :

(+) : bonne performance de l'indicateur,

(#) : performance moyenne de l'indicateur,

(-) : Mauvaise performance de l'indicateur.

Comportements	**Indicateurs**	**Performances**	**Utilisation**
Lecture	La durée de lecture	(+)	Durée complète Durée d'activation
	Les mouvements de la souris	(+) [93] (-) [88] [108]	
	Le nombre de clics de la souris	(+) [88] (#) [93] (-) [108]	
	La durée de défilement du scrollbar	(+) [108]	Combinée avec la durée de défilement (clavier& souris)
	Défilement avec souris	(+) [93] (-) [108]	Combiner avec d'autres indicateurs
	Nombre de clics scrollbar	(+) [88] [69] [44]	
	Défilement avec les touches clavier	(#) [108] [88] [93]	Combiner avec l'indicateur de la souris
	La sélection de texte	(#) [88] [93]	Définir une bonne mesure de distance
Sauvegarde Impression Annotation	Action observée	(-) [94]	Combiner avec d'autres indicateurs

Tableau A.1 — Typologie des comportements & indicateurs d'intérêts associés

B Mise en œuvre de l'outil *Web Cap*

L'outil que nous avons conçu s'inscrit dans le cadre de notre proposition d'un SRI personnalisé. A cet effet, notre premier objectif consiste en la conception et implémentation d'une interface de navigation *web*, baptisée *Web Cap* permettant d'adresser la première partie de notre contribution présentée dans cette thèse, à savoir l'acquisition des données d'interaction pour l'historique de recherche.

Web Cap est un navigateur ayant en plus des fonctionnalités d'un navigateur *web* classique (tel que Internet Explorer ou Mozilla Firefox), la capacité de collecter les documents pertinents pour un utilisateur, lors de ses différentes sessions de recherche et ce, en observant son comportement selon le modèle défini précédemment. Ces documents sont enregistrés dans une base de données (de type MySql). Cette base de données des historiques sera exploitée en perspective par le module de représentation et construction du profil utilisateur.

A ce stade de notre travail, l'outil développé utilise Google comme moteur de recherche et ce, grâce à l'API Google Search plugué au navigateur. Il est néanmoins envisagé par la suite d'intégrer un moteur de recherche personnalisé basé sur le modèle d'accès proposé dans le chapitre 6.

Ainsi, *Web Cap* offre deux modes de fonctionnement : un mode de navigation normal et un mode personnalisé. Par défaut le mode de fonctionnement du navigateur est le mode personnalisé.

B.1 Fonctionnement en mode de navigation normale

En mode normal, *Web Cap* se comporte comme un navigateur *web* classique. Il permet aux utilisateurs de naviguer sur le net en ouvrant des pages *web* simple (HTML) ou contenant du JavaScript, des images, etc. Il permet aussi l'ouverture ou l'enregistrement des fichiers et pages *web*, d'atteindre les pages précédemment visitées et d'actualiser la page courante ou d'arrêter son chargement.

B.2 Fonctionnement en mode de navigation personnalisé

Dans ce mode, *Web Cap* sert d'interface entre l'utilisateur et le moteur de recherche. La figure B.1 présente la page de connexion qui comporte deux boites de dialogue : l'une pour l'administrateur afin qu'il puisse paramétrer et gérer les profils, et la seconde pour les utilisateurs du système afin qu'ils effectuent leurs recherches.

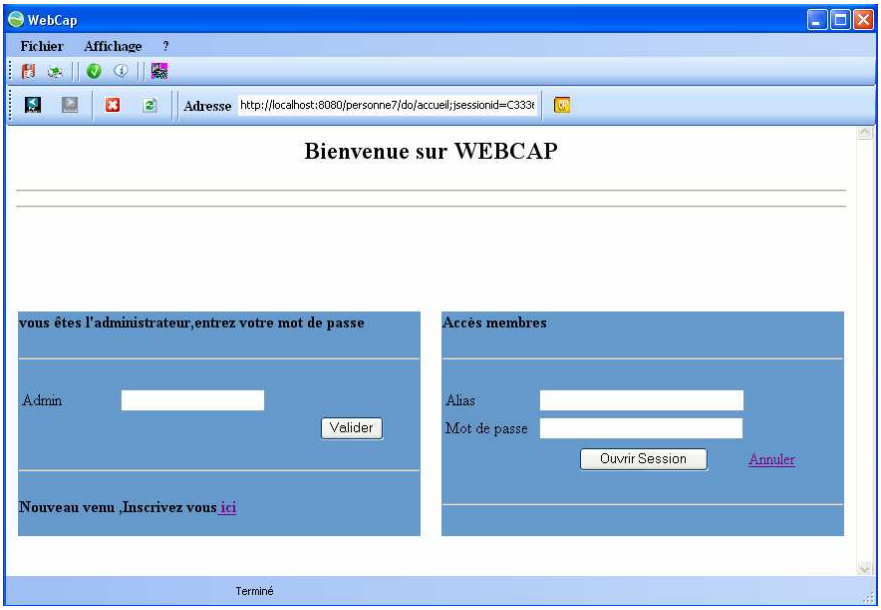


Figure B.1 — Page de connexion de *Web Cap*

B.2.1 Accès utilisateur

A ce niveau on gère des comptes utilisateurs qui correspondent à leur profil. La page d'accueil d'un compte utilisateur est présentée par la figure B.2. Chaque compte contient en plus de ses informations personnelles (nom, prénom, date de naissance, adresse de messagerie, etc.) ses deux dimensions du profil : « *l'historique des interactions* » et « *centres d'intérêt* ». Les centres d'intérêts contenus dans le profil de chaque utilisateur sont obtenus selon l'approche d'inférence décrite précédemment, mais peuvent également être gérés directement par l'utilisateur (ajout, suppression, modification). L'utilisateur peut visualiser à tout moment son historique des interactions.

Lorsqu'un utilisateur soumet une requête au moteur de recherche, des résultats lui sont retournés (voir figure B.3), leur consultation peut alors commencer. Pour chaque document que l'utilisateur consulte,*Web Cap* lui calcule la durée de lecture en secondes, le nombre de clicks de la souris, ses mouvements mesurés en nombre de pixels parcourus ainsi que les mouvements du scroll bar qui sont représentés par la durée en millisecondes du défilement de la page *web* que ce soit avec la souris ou bien avec les touches clavier. Nous tenons à préciser que la durée de lecture commence à être mesurée après le chargement total du document.

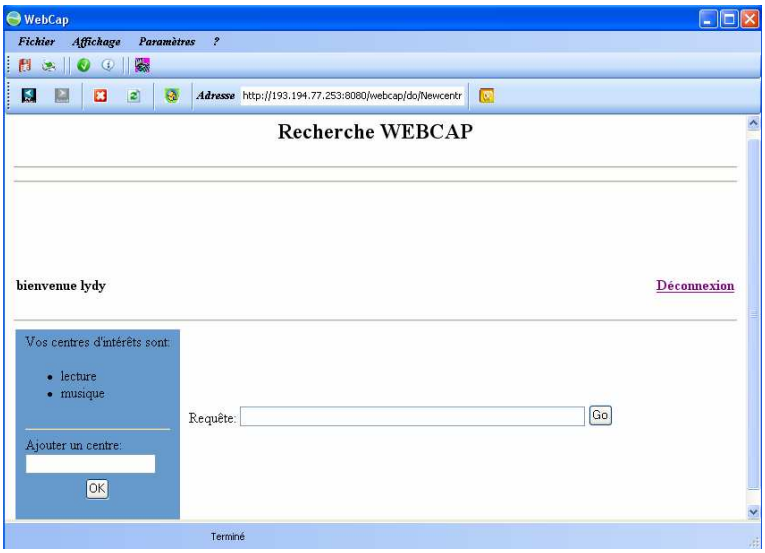


Figure B.2 — Page d'accueil utilisateur

Les valeurs de ces indicateurs seront utilisées pour le calcul du degré d'intérêt de l'utilisateur pour ce document à la fin de la consultation, c'est-à-dire lorsque l'utilisateur effectue l'une de ces tâches de navigation (suivant, précédent ou Fin). A tout moment, l'utilisateur a la possibilité d'évaluer explicitement le document à l'aide du bouton « *évaluer* ». Une fenêtre de valeur de pertinence lui est présentée sur une échelle graduée allant de *très faible* à *très bon* (voir figure B.4).

Il est important à signaler que la déconnexion est obligatoire pour que l'historique de la session en cours soit pris en compte.

Figure B.3 — Exemple de résultats de recherche pour la requête *Java*

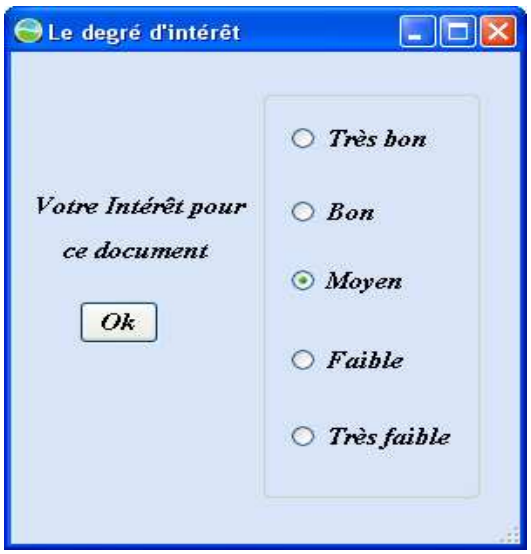


Figure B.4 — Fenêtre des jugements explicites de l'utilisateur

B.2.2 Accès administrateur

Cet accès est exclusivement réservé à l'administrateur du système qui est responsable du paramétrage du navigateur et de la gestion des profils utilisateurs. La figure B.5 montre la page d'accueil réservée à l'administrateur.

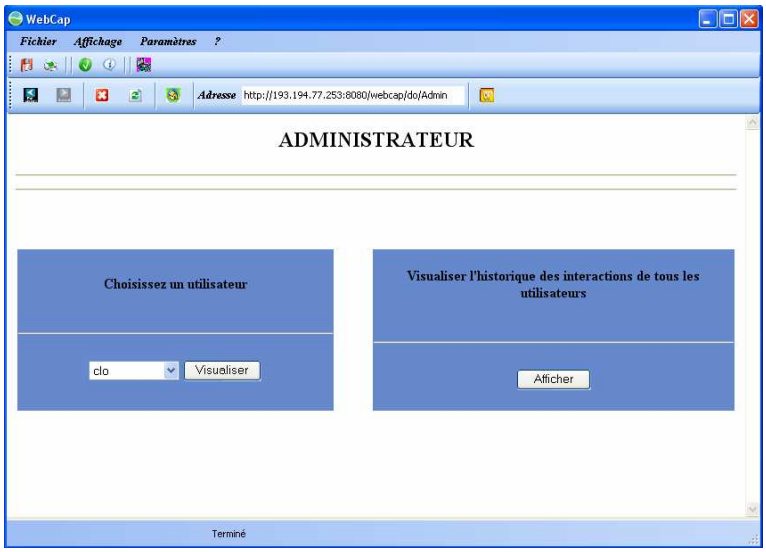


Figure B.5 — Page d'accueil administrateur

Lors de cet accès, l'administrateur peut à ce niveau :

* changer les paramètres de l'application (poids et seuils), et cela à l'aide la fenêtre de paramétrage représentée par la figure B.6,
* visualiser l'historique des interactions d'un utilisateur particulier ou tout les utilisateurs. La figure B.7 montre un exemple de l'historique d'un utilisateur.

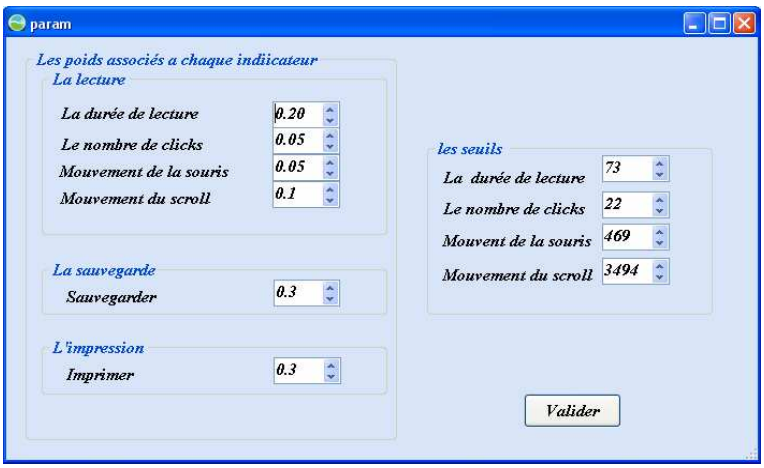


Figure B.6 — Fenêtre de paramétrages des poids et seuils des indicateurs

Figure B.7 — Visualisation de l'historique de recherche d'un utilisateur

C Introduction aux Réseaux Bayésiens

Un réseau Bayésien (RB) (dit « réseau de croyance » ou « réseau probabiliste ») est un modèle graphique orienté sans cycle. Les modèles graphiques sont issus du mariage entre la théorie des graphes et la théorie des probabilités. Un graphe est appréhendé selon un aspect qualitatif et un aspect quantitatif. L'aspect qualitatif est l'ensemble des nœuds du graphe représentant les variables du domaine traité, ainsi que les relations de dépendance entre ces variables. Ces dépendances permettent d'effectuer des inférences, offrant ainsi un support à la prise de décision. L'aspect quantitatif permet d'évaluer les arcs reliant toute paire de nœuds au moyen d'un calcul de probabilités.

C.1 Définition

Soient :

$V = (A, B, ..)$ un ensemble fini de variables ;

$L = V \times V$;

L est l'ensemble des liens reliant une paire de variables de V ; et $G = (V, L)$ est un graphe sur V.

Si le lien est orienté on parle alors d'arcs et G est dit *graphe orienté*. Les variables représentent les événements ou propositions. Une variable peut avoir un nombre donné d'états. Par exemple, une variable peut décrire les couleurs possibles d'un objet, elle peut regrouper des maladies possibles *angine, grippe*. Les états peuvent être discrets ou continus. Les états d'une variable sont mutuellement exclusifs, la couleur d'un jeton ne peut pas être à la fois rouge et noire lorsque les états possibles sont *bleu, rouge, noir, blanc, jaune*. De plus,

- pour tout arc AB, A est l'origine ou le parent de B et B est le nœud final ou le fils de A ;
- un nœud racine est un nœud sans parents ;
- une feuille est un nœud sans fils ;
- un chemin est une séquence de nœuds reliés par des arcs ;
- une chaîne est une séquence de nœuds reliés par des liens ;

– un cycle est un chemin qui a le même nœud initial et final ;
– pour tout nœud $A \in V$ du graphe, $PARENTS_A$ est l'ensemble des parents de A.

C.2 Relations de dépendance

Les réseaux sont utiles pour calculer de façon locale l'impact de la modification d'une information d'une variable sur les états des autres variables. Le changement d'un état d'une variable suite à la réception d'une information dans le réseau dépend de la topologie du graphe, et trois situations principales sont possibles.

C.2.1 Connexion en série

Soit la situation de la figure C.1. A a une influence sur B qui a une influence sur C. L'information peut circuler de A vers C ou de C vers A à travers B dans les deux cas. Par contre, si B est connue ou instanciée, la voie est bloquée, ainsi A et C deviennent indépendants. On dit dans ce cas, que A et C sont *d-séparés* étant donnée B, lorsque B est instanciée.

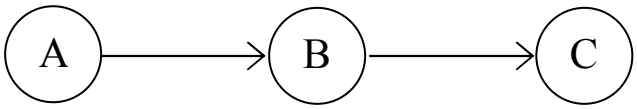


Figure C.1 — Connexion en série

C.2.2 Connexion divergente

L'information peut passer entre les enfants de A lorsque la variable A est non instanciée. Dans la figure C.2, les enfants B, C, D sont dits d-séparés par A.

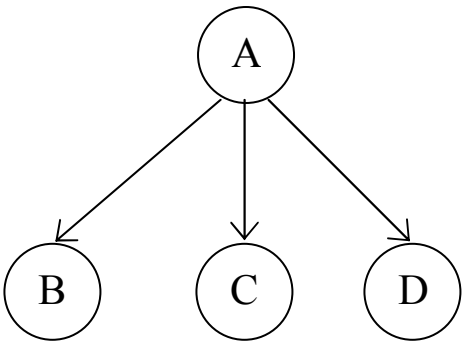


Figure C.2 — Connexion divergente

C.2.3 Connexion convergente

Dans ce type de connexion telle que c'est décrite dans la figure C.3, aucune information n'est donnée sur le nœud fils mise à part l'information apportée par les parents. Les parents sont dits dans ce cas indépendants. Par contre, si l'état du fils est connu alors la cause, c'est-à-dire un des états des parents va pouvoir donner de l'information sur les états des autres parents. L'information peut circuler dans une connexion convergente uniquement lorsque la variable de la connexion ou un de ses descendants a reçu de l'information.

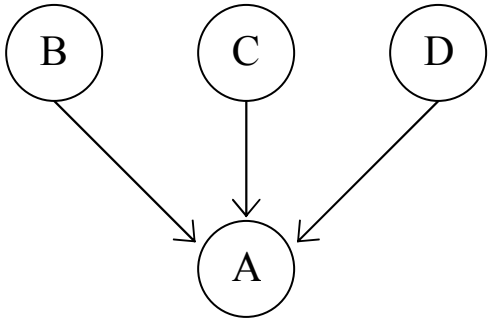


Figure C.3 — Connexion convergente

C.2.4 La d-séparation

Les situations décrites ci-dessus recouvrent les manières possibles de transmettre l'information à travers un réseau. Deux variables distinctes A, B d'un réseau sont d-séparées, si pour tout chemin entre A et B, il existe une variable intermédiaire C, distincte de A et de B telle que :
– soit la connexion est en série ou divergente et C est instanciée,
– ou la connexion est convergente et ni C, ni ses descendants ne sont instanciés.
Ainsi, si deux variables A et B sont d-séparées, alors tout changement d'état dans A n'aura pas d'impact sur l'état de B.

Un réseau Bayésien est doté d'une composante qualitative et d'une composante quantitative. Dans la section qui suit, nous donnons les techniques utilisées pour quantifier les liens existants entre toute paire de nœuds.

C.3 Calcul des probabilités

Les calculs de probabilités permettent de quantifier les liens reliant toute paire de nœuds du réseau. Plusieurs méthodes existent dans la littérature pour quantifier ces liens, nous nous restreignons dans cette partie au calcul Bayésien, à savoir les calculs classiques des probabilités.

C.3.1 Axiomes de base

La probabilité d'un événement A, notée $P(A)$ est un nombre de l'intervalle $[0,1]$. Les probabilités obéissent aux axiomes suivants :

1. $P(A) = 1$ si A est certain ;
2. Si A et B sont mutuellement exclusifs, alors $P(A \cup B) = P(A) + P(B)$.

C.3.2 Probabilités conditionnelles

Les réseaux Bayésiens sont des modèles graphiques probabilistes permettant de représenter les influences entre des événements. Un réseau Bayésien est défini par un graphe acyclique orienté $G = (V,L)$. Dans ce graphe, V représente l'ensemble des nœuds du graphe et L l'ensemble des arcs reliant des paires de nœuds. Chaque nœud V_i représente une variable aléatoire associée à une distribution de probabilités, et chaque arc définit une influence du nœud de départ sur le nœud d'arrivée. La distribution de probabilités associée à une variable spécifie les probabilités de ses états conditionnellement aux états des variables qui l'influencent. On note $P(Parents_{(V_i)})$ où $Parents_{(V_i)}$ représentent l'ensemble des parents de la variable V_i.

Définition.

Soit p une distribution de probabilités jointe sur un ensemble de variables V , et $G = (V,L)$ un graphe acyclique orienté. (G,p) est un réseau Bayésien si chaque variable $A \in V$ est conditionnellement indépendante de ses non descendants, noté $NONPARENTS_A$ étant donné l'ensemble de ses parents, $PARENTS_A$. Pour chaque variable A du graphe, les probabilités conditionnelles suivantes sont définies :

– Si $PARENTS_A = 0$, ce qui signifie que le nœud A est un nœud racine, alors la probabilité a priori de A doit satisfaire :
$$\sum_a P(a) = 1,$$
telle que a constitue l'ensemble des instances possibles de A.
– Si $PARENTS_A \neq 0$ alors la probabilité conditionnelle de A dans le contexte de ses parents est :
$$\sum_a P(a|\theta_A) = 1,$$
où θ_A représente les instances possibles de l'ensemble des parents de A.

C.3.3 La règle de chaînage

Soit $V = (A_1, A_2, .., A_n)$ un ensemble de variables. La probabilité jointe $P(A_1, A_2, .., A_n)$ permet de calculer $P(A_i)$ et $P(A_i/c)$, telle que c est une information donnée. Le nombre et le temps de calcul effectués pour obtenir la probabilité $P(V)$ augmentent d'une manière exponentielle par rapport au nombre de variables contenues dans V.

La règle de chaînage permet de calculer $P(V)$ d'une manière plus rapide lorsqu'il y a des dépendances entre les variables. Ainsi, la probabilité jointe est donnée par :

$$p(V) = \prod_{i=1}^{n} P(A_i | PARENTS_{A_i}) \tag{C.1}$$

où $PARENTS_{A_i}$ constitue l'ensemble des parents de A_i.

Bibliographie

[1] E. Adar and D. Karger. Haystack : Per-user information environments. In *Proceedings of the 8th International Conference on Information Knowledge Management*, pages 413–422, Kansas City, Missouri, November 2-6 1999.

[2] J. Allan, J. Aslam, N. Belkin, C. Buckley, J. Callan, B. Croft, S. Dumais, N. Fuhr, D. Harman, D. Harper, D. Hiemstra, T. Hofmann, E. Hovy, W. Kraaij, J. Lafferty, V. Lavrenko, D. Lewis, L. Liddy, R. Manmatha, A. McCallum, J. Ponte, J. Prager, D. Radev, P. Resnik, S. Robertson, R. Rosenfeld, S. Roukos, M. Sanderson, R. Schwartz, A. Singhal, A. Smeaton, H. Turtle, E. Voorhees, R. Weischedel, J. Xu, and C. Zhai. Challenges in information retrieval and language modeling : report of a workshop held at the center for intelligent information retrieval, university of massachusetts amherst, september 2002. *SIGIR Forum*, 37(1) :31–47, 2003.

[3] Altavista. Altavista search engine.

[4] G. Amato and U. Staraccia. User profile modelling and applications to digital librairies. In *Proceedings of the 3rd European Conference on Research and avanced technology for digital libraries*, pages 184–187, 1999.

[5] R. Armstrong, D. Freitag, D. Joachims, and T. Mitchell. Webwatcher : A learning apprentice for the world wide web. In *Spring symposium on Information gathering from Heterogeneous, distributed environments*, pages 6–12, 1995.

[6] F. Asnicar and C. Tasso. ifweb : A prototype of user model-based intelligent agent for documentation filtering and navigation in the world wide web. In *Proceedings of the 6th International Conference on User Modeling*, pages 3–11, Chia Laguna, Sardinia, Italy, June 2-5 1997.

[7] R. Baeza-Yates and R. A. Ribeiro-Neto. *Modern Information Retrieval*. New York : ACM Press ; Harlow England : Addison-Wesley, cop., 1999.

[8] M. Balabanovic and Y. Shoham. Fab : Content-based collaborative recommendation. *Communications of the ACM*, 3(40) :66–72, March 1997.

[9] R. Barrett, P. Maglio, and D. Kellem. How to personalize the web. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 75–82, New York, NY, USA, 1997. ACM.

[10] M. Bates. Search techniques. In *Annual Review of Information Science and Technology 16*, pages 139–169. M.E. Williams, ed., 1981.

[11] M. Baziz. Towards a semantic representation of documents by ontology-document mapping. In *The Eleventh International Conference on Artificial Intelligence*, pages 33–43. Christoph Bussler, Diete Fensel (Eds.), LNCS/LNAI 3192, Springer, Springer-Verlag Berlin, Heidelberg, Germany, 2-4 septembre 2004.

[12] R. Belew. Adaptive information retrieval : Using a connectionist representation to retrieve and learn about documents. In *Proceedings of the 12th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 11–20, Cambridge, Massachusetts, United States, 1989. ACM.

[13] N. Belkin, C. Cool, D. Kelly, S.-J. Lin, S. Y. Park, J. Perez-Carballo, and C. Sikora. Iterative exploration, design and evaluation of support of query reformulation in interactive information retrieval. *Information Processing & Management*, 37(3), 2001.

[14] N. Belkin and W. Croft. Information filtering and information retrieval : Two sides of the same coin ? *Communication of the ACM*, 35(12) :29–38, 1992.

[15] D. Billsus and M. Pazzani. A hybrid user model for news stories classification. In *Proceedings of the seventh International Conference on User Modelins*, pages 99–108, Banff, Canada, 1999.

[16] D. C. Blair and M. E. Maron. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Commun. ACM*, 28(3) :289–299, 1985.

[17] E. Bloedorn, I. Mani, and T. MacMillan. Machine learning of user profiles : Representational issues. In *Proceedings of of the Thirteenth National Conference on Artificial Intelligence*, pages 433–438, Portland, Oregon, August 1996.

[18] P. Borlund. The iir evaluation model : a framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3) :152–179, Avril 2003.

[19] P. Borlund and P. Ingwersen. Measures of relative relevance and ranked half-life : performance indicators in interactive ir. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 324–331, Melbourne, Australia, August 1998. Croft W.B et al Eds, ACM.

[20] J. Bottraud, G. Bisson, and M. Bruandet. Expansion de requêtes par apprentissage automatique dans un assistant pour la recherche d'information. In *Actes du congrès CORIA*, pages 89–105, Mars 2004.

[21] M. Boughanem. *Systèmes de recherche d'informations : d'un modèle classique à un modèle connexionniste*. Thèse de doctorat, Université Paul Sabatier de Toulouse, 1992.

[22] M. Boughanem, A. Brini, and D. Dubois. Possibilistic networks for information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference. Workshop Information retrieval and applications of graphical models*, Amsterdam, juillet 2007. ACM DL.

[23] M. Boughanem, C. Chrisment, and C. Soulé-Dupuy. Query modification based on relevance back-propagation in an ad hoc environment. *Information processing & management*, 35(2) :121–139, 1999.

[24] M. Boughanem, C. Chrisment, and L. Tamine. On using genetic algorithms for multimodal relevance optimisation in information retrieval. *Journal of American Society in Information Systems*, 53(11) :934–942, 2002.

[25] M. Boughanem and C. Soulé-Dupuy. Mercure at trec-6. In *6th International Conference on Text Retrieval*, pages 321–328, Washington, USA, novembre 1997. E. M. Voorhees and D. K. Harman Editors.

[26] M. Boughanem, H. Tebri, and M. Tmar. Irit at trec 2002 : Filtering track. In *The Eleventh Text Retrieval Conference (TREC 2002)*, Gaithersburg, Maryland(USA), novembre 2002. Text Retrieval Conference, TREC.

[27] C. Bourne and B. Anderson. Dialog labworkbook. PaloAlto, Californie, USA, 1979. Second edition, Looked Information Systems.

[28] M. Bouzeghoub and D. Kostadinov. Personnalisation de l'information : Aperçu de l'état de l'art et définition d'un modèle flexible de définition de profils. In *Actes de la seconde édition de la Conférence en Recherche d'Information et Applications (CORIA)*, pages 201–218, Grenoble, France, 2005.

[29] C. Bradford and I. Marshall. Analysing users www search behaviour. *Lost in the Web - Navigation on the Internet, IEE Colloquium*, 6(169) :1–4, 1999.

[30] G. Brajnik, S. Mizzaro, , and C. Tasso. Evaluating user interfaces to information retrieval systems : a case study on user support. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 128–136, Zurich, 1996.

[31] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7) :107–117, 1998.

[32] A. Brini. *Application des réseaux bayésiens possibiliste à la recherche d'information*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France, décembre 2005.

[33] A. Brini, M. Boughanem, and D. Dubois. A model for information retrieval based on possibilistic networks. In *String Processing and Information Retrieval*, pages 271–282, Buenos Aires, ARGENTINE, janvier 2005. LNCS, Springer Verlag.

[34] G. Cabanac, M. Chevalier, C. Chrisment, and C. Julien. Collective annotation : Perspectives for information retrieval improvement. In *Large-Scale Semantic Access to Content (Text, Image, Video and Sound) (RIAO), Pittsburgh, USA, 30/05/07-01/06/07*, page (electronic medium), http ://www.le-cid.org, mai 2007. Centre de hautes études internationales d'Informatique Documentaire (C.I.D.).

[35] J. Chaffee and S. Gauch. Personal ontologies for web navigation. In *Proceedings of the ninth international conference on Information and knowledge management*, pages 227–234, McLean, Virginia, United States, 2000. ACM Press.

[36] V. Challam. Contextual information retrieval using ontology based user profiles. Master's thesis, Jawaharlal Nehru Technological University, 2004.

[37] P. Chan. A non invasive learning approach to building web user profiles. In *Workshop on Web usage analysis and user profiling, Fifth International Conference on Knowledge Discovery and Data Mining*, pages 7–12, San Diego, August 1999.

[38] C. Chen, M. Chen, and Y. Sun. Pva : A self-adaptive personal view agent. *Journal of Intelligent Information Systems*, 18(2-3) :173–194, Mars 2002.

[39] L. Chen and K. Sycara. Webmate : A personal agent for browsing and searching. In *Proceedings of the 2nd international conference on autonomous agents and multi agent systems, Minneapolis*, pages 10–13, 1998.

[40] W. Chien. Learning query behavior in the haystack system. Master's thesis, MIT, USA, June 2000.

[41] P. A. Chirita, D. Olmedilla, and W. Nejdl. Pros : a personalized ranking platform for web search. In *Proceedings of the 3rd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, pages 34–43, Eindhoven, Netherlands, August 2004.

[42] K. Church and P. Hanks. Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1) :22–29, Mars 1990.

[43] V. Claveau and P. Sébillot. Extension de requêtes par lien sémantique nom-verbe acquis sur corpus. In *Actes de la 11ème conférence de Traitement automatique des langues naturelles*, Fês, Maroc, avril 2004.

[44] M. Claypool, P. Le, M. Wased, and D. Brown. Implicit interest indicators. In *Proceedings of the 6th ACM International Conference on Intelligent User Interfaces*, pages 33–40, Santa Fe, New Mexico, USA, Janvier 2001.

[45] C. Cleverdon. The cranfield test on index language devices. *Aslib*, 19(6) :173–194, 1967.

[46] C. Cool and A. Spink. Issues of context in information retrieval : an introduction to the special issue. *Journal of Information Processsing and Management*, 38(55) :605–611, 2002.

[47] F. Crestani. An adaptive information retrieval system based on neural networks. In *Proceedings of the International Workshop on Artificial Neural Networks (IWANN)*, pages 732–737, London, UK, 1993. Springer-Verlag.

[48] F. Crestani, M. Lalmas, C. J. Van Rijsbergen, and I. Campbell. Is this document relevant ? ... probably. *A Survey of Probabilistic Models In Information Retrieval, ACM Computing Surveys*, 30(4), December 1998.

[49] F. Crestani and I. Ruthven. Introduction to special issue on contextual information retrieval systems. *Information Retrieval*, 10(2) :111–113, 2007.

[50] W. Croft, R. Cook, and D. Wilder. Providing government information on the internet : Experiences with thomas. In *Proceedings of the Second Annual Conference on the Theory and Practice of Digital Libraries (DL'95)*, pages 19–24, Austin, TX, Juin 1995.

[51] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma. Query expansion by mining user logs. *IEEE Transactions on Knowledge and Data Engineering*, 15(4) :829–839, Jul/Aug 2003.

[52] S. Deerwster, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society Information Science and Technology*, 41(6) :391–407, 1990.

[53] P. Denning. Acm president's letter : electronic junk. *Communication of ACM*, 25(3) :163–165, 1982.

[54] N. Denos, C. Berrut, L. Gallardo-Lopez, and A. Nguyen. Proceedings of the cocofil : une plateforme de filtrage collaboratif orientée vers la communauté. In *Proceedings of the 1st COnférence en Recherche d'Information et Applications.*, pages 9–26, 2004.

[55] G. Desktop. The web organized by topic into categories.

[56] P. Dourish. What we talk about when we talk about context. *Personal Ubiquitous Computer*, 8(1) :19–30, 2004.

[57] S. Dumais, E. Cuttrel, J. Cadiz, G. Jancke, R. Sarin, and D. Robbins. Stuff i've seen : a system for a personal information retrieval and re-use. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development*, pages 72–79, Toronto, Canada, 2003. ACM Press.

[58] Excite. http ://www.excite.fr/.

[59] R. Fidel. Searchers' selection of search keys. *Journal of the American Society of Information Science (JASIS)*, 34, 1991.

[60] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. Placing search in context : the concept revisited. In *Proceedings of the 8th International World Wide Web Conference*, pages 406–414, 2001.

[61] C. Fluhr and F. Debili. Interrogation en langue naturelle de données textuelles et factuelles. In *Intelligent Multimédia Information System and Management*, Grenoble, France, 1985.

[62] P. W. Foltz. Using latent semantic indexing for information filtering. *SIGOIS Bull.*, 11(2-3) :40–47, 1990.

[63] M. E. Frisse and S. B. Cousins. Information retrieval from hypertext : update on the dynamic medical handbook project. In *Proceedings of the 2sd Annual ACM Conference on Hypertext*, pages 199–212, New York, NY, USA, 1989. ACM Press.

[64] N. Fuhr. Information retrieval : introduction and survey. post-graduate course on information retrieval, university of Duisburg-Essen, Germany, 2000.

[65] S. Gauch, J. Chaffe, and A. Pretschner. Ontology-based personalized search and browsing. *Web Intelligence and Agent System*, 1(3,4) :219–234, 2003.

[66] S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli. User profiles for personalized information access. *The Adaptive Web*, 4321 :54–89, 2007.

[67] G. Gentili, A. Micarelli, and F. Sciarrone. Infoweb : An adaptive information filtering system for the cultural heritage domain. *Applied Artificial Intelligence*, 17(8-9) :715–744, 2003.

[68] D. Godoy. *Learning User Interests for User Profiling in Personal Information Agents*. PhD thesis, Departamento de Computación y Sistemas, Facultad de Ciencias Exactas Univ. Nacional del Centro de la Provincia de Buenos Aires, Tandil, Bs. As., Argentina, 2005.

[69] J. Goecks and J. Shavlik. Learning user's interests by unobtrusively observing their normal behaviour. In *Proceedings of the 5th international conference on Intelligent user interfaces*, pages 129–132, New Orleans,LA, 2000. ACM.

[70] N. Guarino, C. Masolo, and G. Vetere. Ontoseek : Content-based access to the web. *IEEE Intelligent Systems*, 14(3) :70–80, Mai 1999.

[71] D. Harman. Relevance feedback revisited. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development*, pages 1–10. N. Belkin, P. Ingwersen, A. Mark Pejtersen Eds, 1992.

[72] T. H. Haveliwala. Topic-sensitive pagerank. In *Proceedings of the Eleventh International World Wide Web Conference (WWW 2002)*, Honolulu, Hawaii, May 2002.

[73] N. Hernandez, J. Mothe, C. Chrisment, and D. Egret. Modeling context through domain ontologies. *Journal of Information Retrieval, Contextual Information Retrieval Systems*, 10(2) :143–172, avril 2007.

[74] T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of Uncertainty in Artificial Intelligence, UAI'99*, Stockholm, 1999.

[75] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1-2) :177–196, 2001.

[76] I. Hsieh-Yee. Effects of the searcher and experience and subject knowledge on the search tactics of novice and exprienced searchers. *Journal of the American Society for Information Science*, 44(3) :161–174, 1993.

[77] B. Hugues and N. Jian-Yun. Modèles de langue appliqués à la recherche d'information contextuelle. In *Conférence en recherche information et applications CORIA'2006*, pages 213–224, Lyon , FRANCE, 2006.

[78] M. N. Huhn and L. Stephens. Personal ontologies. *Internet Computing*, 3(5), October 1999.

[79] P. Ingwersen. Cognitive perspectives of information interactions : Elements of a cognitive ir theory. *Annual review of information science and technology*, 52(1) :3–50, 1996.

[80] P. Ingwersen and K. Järvelin. *The TURN : Integration of Information Seeking and Retrieval in Context*. SPRINGER, August 2005.

[81] J. Janes. Relevance judgements and the incremental presentation of document representation. *Information Processing & Management*, 27(6) :629–646, 1991.

[82] K. Jarvelin and J. Kekalainen. Ir evaluation methods for highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 41–48. Belkin and al Eds, 2000.

[83] K. Jarvelin and J. Kekalainen. Cumulative gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4) :422–446, 2002.

[84] G. Jeh and J. Widom. Scaling personalized web search. In *Proceedings of the 12th international conference on World Wide Web*, pages 271–279, New York, NY, USA, 2002. ACM Press.

[85] F. Jensen. *Bayesian networks and decision graphs*. Springer, 2001.

[86] X. Jin, Y. Zhou, and B. Mobasher. Web usage mining based on probabilistic latent semantic analysis. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'04)*, pages 197–205, Seattle WA, USA, 2004. ACM.

[87] Y. Jing and W. Croft. An association thesaurus for information retrieval. In *Proceedings of the 4th International Conference Recherche d'Information Assistee par Ordinateur*, pages 146–160, New York, US, 1994.

[88] K. Jung. Modeling web user interest with implicit indicators, master thesis. Master's thesis, Florida Institute of Technology, 2001.

[89] I. Kang and G. Kim. Integration of multiple evidences based on a query type for web search. *Information Processing & Management*, 40(3) :459–478, 2004.

[90] D. Kelly and N. J. Belkin. Reading time, scrolling and interaction : Exploring implicit sources of user preferences for relevance feedback during interactive information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 408–409, New Orleans, Louisiana, United States, 2001.

[91] D. Kelly and J. Teevan. Implicit feedback for inferring user preference : a bibliography. *SIGIR Forum*, 37(2) :18–28, 2003.

[92] N. J. Kelly. Understanding implicit feedback and document preference : a naturalistic study. In *PHD dissertation*. Ritgers University, New Jersey, January 2004.

[93] H. Kim and P. Chan. Implicite indicators for interstiong web pages. Technical Report CS-2005-05, Department of Computer Sciences, Florida, Institute of Technology, Melbourne, FL. 32901, USA, 2005.

[94] H. Kim and P. K. Chan. Learning implicit user interest hierarchy for context in personalization. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 101–108, Miami, Florida, USA, 2003. ACM Press.

[95] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 5(46) :604–632, 1999.

[96] K. Knight and S. Luk. Building a large knowledge base for machine translation. In *Proceedings of American Association of Artificial Intelligence Conference*, pages 773–778, Orlando, Florida, July 18-22 1999.

[97] M. Kobayashi and K. Takeda. Information retrieval on the web. *ACM Computing Surveys*, 32(2) :144–173, 2000.

[98] A. Kobsa. Generic user modeling systems. *User Modeling and User Adapted Interaction Journal*, (11) :49–63, 2001.

[99] A. Kobsa. Privacy-enhanced web personalization. In *The Adaptive Web : Methods and Strategies of Web Personalization, Lecture Notes in Computer Science*, volume 4321, Berlin Heidelberg New York, 2007. In Brusilovsky, Kobsa, P. and Nejdl, A. W. (eds.), Springer-Verlag.

[100] A. Kobsa and W. Wahlster. *User Models in Dialog Systems*. Springer-Verlag, 1989.

[101] J. Konstan, M. Miller, B. ans Maltz, J. Herlocker, L. Gordon, and J. Riedl. Grouplens : Applying collaborative filtering to usenet news. In *Communications of the ACM, 40(3)*, pages 77–87, 1997.

[102] D. Kostadinov. La personnalisation de l'information,définition de modèle de profil utilisateur. rapport de dea. Master's thesis, Université de Versailles, France, 2003.

[103] D. Kostadinov. *Personnalisation de l'information : une approche de gestion de profils et de reformulation de requêtes*. PhD thesis, L'UNIVERSITE DE VERSAILLES SAINT-QUENTIN-EN-YVELINES, Décembre 2007.

[104] G. Koutrika and Y. Ioannidis. Personalization of queries in database systems. In *Proceedings of the 20th International Conference on Data Engineering*, page 597, Washington, DC, USA, 2004. IEEE Computer Society.

[105] B. Krovetz. Homonymy and polysemy in information retrieval. In P. R. Cohen and W. Wahlster, editors, *Proceedings of the e COLING/ACL'97*, pages 72–79, Somerset, New Jersey, 1997. Association for Computational Linguistics.

[106] K. Kwok. A neural network for probabilistic information retrieval. *SIGIR Forum*, 23(SI) :21–30, 1989.

[107] Y. Labrou and T. Finin. Yahoo ! as an ontology - using yahoo ! categories to describe documents. In *Proceedings of the 8th International Conference On Information Knowledge Management*, pages 180–187, Kansas City, Missouri, November 2-6 1999.

[108] P. Le and M. Waseda. A curious browser : Implicit ratings. Technical report, 2000.

[109] A. Lelu and C. François. Automatic generation of hypertext links in information retrieval systems. In *Communication of colloque ECHT'92*, New York, 1992. ACM Press.

[110] H. Lieberman. Letizia : An agent that assists web browsing. In *Proceedings of the 14th International Joint Conference On Artificial Intelligence*, Montreal, Canada, August 1995.

[111] C. Lin, G. Xue, H. Zeng, and Y. YU. Using probabilistic latent semantic analysis for personalized web search. In *Proceedings of the APWeb Conference*, pages 707–717, Berlin Heidelberg, 2005. Springer-Verlag.

[112] S. Lin, C. Shih, M. Chen, J. Ho, M. Ko, and Y. M. Huang. Extracting classification knowledge of internet documents with mining term-associations : A semantic approach. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 241–249, Melbourne, Australia, 1998.

[113] F. Liu and C. Yu. Personalized web search for improving retrieval effectiveness. *IEEE Transactions on knowledge Data Engineering*, 16(1) :28–40, 2004.

[114] F. Liu, C. Yu, and W. Meng. Personalized web search by mapping user queries to categories. In *Proceedings of the 11th International Conference on Information and Knowledge Management*, pages 558–565, Mclean, Virginia, November 4-9 2002. ACM.

[115] L. B. Lorigo, H. Pan, T. Hembrooke, L. Joachims, and G. Granka. The influence of task and gender on search and evaluation behavior using google. *Information Processing & Management*, 42 :1123–1131, 2006.

[116] H. P. Luhn. A statistical approach to mechanized encoding and searching of literary information. *jounal of IBM-JRD*, 1(4) :309–317, october 1957.

[117] T. Malone, K. Grant, F. Turbak, S. Brobst, and M. Cohen. Intelligent information-sharing systems. *Communication of ACM*, 30(5) :390–402, 1987.

[118] H. Marais and K. Bharat. Supporting cooperative and personal surfing with a desktop assistant. In *roceedings of the 10th annual ACM symposium on User interface software and technology*, pages 129–138, Banff, Alberta, Canada, October 14-17 1997.

[119] J. Mc Gowan. A multiple model approach to personnalised information access. Master's thesis, Faculty of science, University college Dublin, February 2003.

[120] A. Micarelli and F. Sciarrone. Anatomy and empirical evaluation of an adaptive web-based information filtering system. *User Modeling and User-Adapted Interaction*, 14(2-3) :159–200, June 2004.

[121] S. Middleton, N. Shadbolt, and D. De Roure. Capturing interest through inference and visualization : Ontological user profiling in recommender systems. In *International Conference on Knowledge Capture, K-CAP*, pages 62–69, Sanibel Island, Florida, September 2003.

[122] G. Miller. Wordnet : a lexical database for english. *Commun. ACM*, 38(11) :39–41, 1995.

[123] M. Minio and C. Tasso. User modeling for information filtering on internet services : Exploiting an extended version of the umt shell. In *Workshop on User Modeling for Information Filtering on the WWW*, Kailua-Kona, Hawaii, January 2-5 1996.

[124] T. Mitchell. *Machine Learning*. McGraw-Hill Higher Education, 1997.

[125] D. Mladenic. Personal webwatcher : design and implementation. In *Technical Report IJS-DP-7472*. J. Stefan Institute, Department for Intelligent Systems, 1998.

[126] B. Mobasher. *Data Mining for Web Personalization*, volume 4321. The Adaptive Web : Methods and Strategies of Web Personalization, Berlin Heidelberg New York, springer-verlag edition, 2007. Lecture Notes in Computer Science.

[127] B. Mobasher, T. Dai, H. andLuo, Y. Sun, and J. Zhu. Integrating web usage and content mining for more effective personalization. In *In E-Commerce and Web Technologies : Proceedings of the EC-WEB Conference*, pages 165–176. Lecture Notes in Computer Science (LNCS) 1875, Springer, 2000.

[128] M. Montebello, W. Gray, and S. Hurley. A personal evolvable advisor for www knowledge-based systems. In *Proceedings of the 1998 International Database Engineering and Application Symposium*, pages 224–233, Cardiff, Wales, U.K, July 1998.

[129] M. Morita and Y. Shinoda. Information filtering based on user behavior analysis and bestmatch text retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 272–281, Dublin, Ireland, 1994.

[130] A. Moukas. Amalthaea : Information discovery and filtering using a multiagent evolving ecosystem. *Applied Artificial Intelligence*, 11(5) :437–457, 1997.

[131] N. Nanas, U. Uren, and A. Deroeck. Building and applying a concept hierarchy representation of a user profile. In *Proceedings of the 26th Annual ACM Conference on Research and Development in Information Retrieval SIGIR*, pages 154–204, 2003.

[132] D. Nichols. Implicit rating and filtering. In *In Proc. 5th DELOSWorkshop on Filtering and Collaborative Filtering*, pages 31–36, Budapest, Hungary, November 1997.

[133] D. W. Oard and J. Kim. Modeling information content using observable behavior. In *In Proceedings of the 64th Annual Meeting of the American Society for Information Science and Technology*, pages 38–45, USA, 2001.

[134] M. J. Pazzani, J. Muramatsu, and D. Billsus. Syskill & webert : Identifying interesting web sites. In *Proceedings of the 30th National Conference on Artificial Intelligence*, pages 54–61, Portland, 1996.

[135] J. Pearl. *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Morgan Kaufmann publishers Inc., San Francisco, CA, USA, 1988.

[136] J. Pitkow, H. Schütze, and C. T. et all. Personalized search. *Communications of the ACM*, 9(45) :50–55, 2002.

[137] W. Pohl. *Logic-Based Representation and Inference for User Modeling Shell Systems*. PhD thesis, 1997.

[138] A. Pretschner. Ontology based personalized search. Master's thesis, University of Kansas, June 1999.

[139] Y. Qiu and H. Frei. Concept based query expansion. In *Proc. 16th Int'l ACM SIGIR Conf. R & D in Information Retrieval*, pages 160–169, 1993.

[140] L. Quiroga and J. Mostafa. Empirical evaluation of explicit versus implicit acquisition of user profiles in information filtering systems. In *Proceedings of the 63rd annual meeting of the American Society for Information Science and Technology*, pages 4–13, Medford, NJ, 2000. D. H. Kraft (Ed.).

[141] L. Razmerita. *Modèle Utilisateur et Modélisation Utilisateur dans les Systèmes de Gestion des connaissances : une approche fondée sur les ontologies*. PhD thesis, Université Paul Sabatier, Toulouse III, Décembre 2003.

[142] S. Robertson and K. Sparck Jones. Relevance weighting for search terms. *Journal of The American Society for Information Science*, 27(3) :129–146, 1976.

[143] S. Robertson, S. Walker, and M. Beaulieu. Okapi at trec-7 : Automatic ad hoc, filtering, vlc and interactive track, 1999.

[144] S. Robertson, S. Walker, and M. Beaulieu. Experimentation as a way of life : Okapi at trec. *Information Processing & Management*, 36(1) :95–108, 2000.

[145] S. Robertson, S. Walker, M. Sparck Jones, and al. Okapi at trec-3. In *Second Text Retrieval Conf (TREC-3)*, pages 109–26, 1995.

[146] J. Rocchio. Relevance feedback in information retrieval. In *The SMART retrieval system - experiments in automatic document processing*, pages 313–323, Englewood Cliffs, 1971. Prentice Hall.

[147] I. Ruthven and M. Lalmas. A survey on the use of relevance feedback for information access systems. *Knowledge Engeneering review*, 18(2) :95–145, 2003.

[148] H. Sakagami and T. Kamba. Learning personal preferences on online newspaper articles from user behaviors. In *Proceedings of the 6th International WWW Conference*, Santa Clara, California, April, 7-11 1997.

[149] G. Salton. *The SMART Retrieval System—Experiments in Automatic Document Processing*. Prentice-Hall Inc, NJ, 1971.

[150] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Jounal of American Society for Information Science*, 41(4) :288–297, 1990.

[151] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

[152] G. Salton and R. K. Waldstein. Term relevance weights in on-line information retrieval. *Information Process & Management.*, 14(1) :29–35, 1978.

[153] G. Salton and C. Yang. On the specification of term values in automatic indexing. *Journal of documentation*, pages 351–372, 1973.

[154] G. Saporta. *Probabilités, analyse de données et statistique*. Editions technip edition, 1990.

[155] T. Saracevic. The stratefied model of information retrieval interaction : extension and applications. In *Proceedings of the 60th annual meeting of the American Society for Information Science*, pages 313–327, Medford, NJ, 1997.

[156] Seruku. Toolbar.

[157] J. Shavlik, S. Calcari, T. Eliassi-Rad, and J. Solock. An instructable, adaptive interface for discovering and monitoring information on the world wide web. In *Proceedings of the 1999 International Conference on Intelligent User Interfaces*, pages 157–160, Redondo Beach, California, January 5-8 1999.

[158] J. Shavlik and T. Eliassi-Rad. Intelligent agents for web-based tasks : An advice-taking approach. In *Working Notes of the AAAI/ICML-98 Workshop on Learning for text categorization*, Madison, WI, July 26-27 1998.

[159] X. Shen, B. Tan, and C. Zhai. Context-sensitive information retrieval using implicit feedback. In *In Proceedings of the 29th annual internationsl ACM SIGIR Conference on Research and development in Information retrieval*, pages 43–50, Salvador, Brazil, 2005.

[160] A. Sieg, B. Mobasher, and R. Burke. Inferring users information context : Integrating user profiles and concept hierarchies. In *Meeting of the International Federation of Classification Societies, IFCS,*, Chicago, July 2004.

[161] A. Singhal. *Term weighting revisited*. PhD thesis, PHD of Cornell University, 1997.

[162] G. Somlo and A. Howe. Using web helper agent profiles in query generation. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 812–818, New York, NY, USA, 2003. ACM Press.

[163] K. Sparck Jones. *Automatic Keyword Classification for Information Retrieval*. Automatic keyword CL, London, 1971.

[164] K. Sparck Jones. Experiments in relevant weighting of search terms. *Information Processing & Management*, 15(3) :133–144, 1979.

[165] S. Speretta and S. Gauch. Personalizing search based user search histories. In *Proceedings of the 13th International Conference on Information Knowledge and Management*, pages 238–239, 2004.

[166] S. Speretta and S. Gauch. Personalized search based on user search histories. In *Web Intelligence*, pages 622–628, France, 2005. IEEE Computer Society.

[167] C. Starr and P. Shi. An introduction to bayesian belief networks and their applications to land operations. Technical report, Defence Science and Technology Organisation, Edinburgh South Australia, 2004.

[168] A. Stefani and C. Strappavara. Personalizing access to web sites : The siteif project. In *Proceedings of the 2nd Workshop on Adaptive Hypertext and Hypermedia*, Pittsburgh, June 20-24 1998.

[169] C. Stevens. *Knowledge-based assistance for accessing large, poorly structured information spaces*. PhD thesis, University of Colorado, Department of Computer Science, 1993.

[170] SurfSaver. http ://www.surfsaver.com/.

[171] L. Tamine. *Optimisation de requêtes dans un système de recherche d'information : approche basée sur l'exploitation de techniques avancées de l'algorithmique génétique*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France, décembre 2000.

[172] L. Tamine, M. Boughanem, and W. N. Zemirli. Inferring the user's interests using the search history. In M. Schaaf and K.-D. Althoff, editors, *Workshop on information retrieval, Learning, Knowledge and Adaptatbility*, pages 108–110, Hildesheim, Germany, 2006.

[173] L. Tamine, M. Boughanem, and W. N. Zemirli. Exploiting multi-evidence from multiple user's interests to personalizing information retrieval. In *IEEE International Conference on Digital Information Management (ICDIM 2007)*, pages 7–12, Lyon, France, octobre 2007. Youakim Badr, Richard Chbeir, Pit Pichappan (Eds.), IEEE Engineering Management Society.

[174] L. Tamine and S. Calabretto. *Recherche d'information sur le web : état des lieux et perspectives*, chapter Recherche d'information contextuelle et *web*. Number 7. Editions Hermes, sous la direction de Mohand Boughanem et Jacques Savoy, 2008.

[175] L. Tamine, C. Chrisment, and M. Boughanem. Multiple query evaluation based on an enchanced genetic algorithm. *Information Processing & Management*, 39(2) :215–231, 2003.

[176] L. Tamine, W. N. Zemirli, and W. Bahsoun. Approche statistique pour la définition du profil d'un utilisateur de système de recherche d'information. *Information - Interaction - Intelligence*, 7(1), 2007.

[177] A. Tan and C. Teo. Learning user profiles for personalized information dissemination. In *In : Proceedings of 1998 IEEE International Joint Conference on Neural Networks*, pages 183–188, Alaska, May 4-9 1998.

[178] F. Tanudjaja and L. Mui. Persona : A contextualized and personalized web search. In *Proc 35th Hawaii International Conference on System Sciences*, page 53, Big Island, Hawaii, January 2002.

[179] H. Tebri, M. Boughanem, and C. Chrisment. Incremental profile learning based on a reinforcement method. In *Proceedings of the 2005 ACM symposium on Applied computing*, pages 1096–1101, 2005.

[180] J. Teevan, S. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 449–456, Salvador, Brazil, August 15-19 2005.

[181] J. Trajkova and S. Gauch. Improving ontology-based user profiles. In *Proceedings of the 8th Conference of Recherche d'Information Assistée par Ordinateur*, pages 380–389, University of Avignon (Vaucluse), France, April 26-28 2004.

[182] M. Turpeinen and T. Saari. System architecture for psychological customization of communication technology. In *Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences - Track 7*, page 70202.2, Washington, DC, USA, 2004. IEEE Computer Society.

[183] H. Turtle. *Inference networks for document retrieval*. PhD thesis, University of Massachusetts, 1991.

[184] H. Turtle and W. B. Croft. Inference networks for document retrieval. In *Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 1–24, Brussels, Belgium, 1990. ACM Press.

[185] P. Vakkari. Relevance and contributing information types of searched documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 2–9, Athens, Greece, 2000. ACM Press.

[186] P. Vakkari. A theory of the task-based information retrieval process : a summary and generalisation of a longitudinal study. *Journal of documentation*, 57(1) :44–60, 2001.

[187] C. Van Rijsbergen. A non-classical logic for information retrieval. *The computer journal*, 29(6) :481–485, 1986.

[188] E. Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 61–69, Dublin, Ireland, 1994. Springer-Verlag New York, Inc.

[189] W3C. 2005.

[190] J. Wen, N. Lao, and W. Y. Ma. Probabilistic model for contextual retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 57–63, Sheffield, United Kingdom, August 2004.

[191] R. W. White, J. M. Jose, and I. Ruthven. Comparing explicit and implicit feedback techniques for web retrieval : Trec-10 interactive track report. In *Proceedings of the Tenth Text Retrieval Conference*, pages 534–538, Gaithersburg, MD, 2001.

[192] R. W. White, I. Ruthven, and J. M. Jose. The use of implicit evidence for relevance feedback in web retrieval. In *proceedings of 24th BCS-IRSG European Colloquium on IR*, pages 93–109, London, UK, 2002. Springer-Verlag.

[193] R. W. White, I. Ruthven, and J. M. Jose. A study of factors affecting the utility of implicit relevance feedback. In *Proceedings of the 28 th annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 15–19. Marchionini, G. Moffat, A Tait, J Baeza-Yates, R Ziviani, N Eds, August 2003.

[194] D. Widyantoro, J. Yin, M. El Nasr, L. Yang, A. Zacchi, and J. Yen. Alipes : A swift messenger in cyberspace. In *Proceedings of Spring Symposium Workshop on Intelligent Agents in Cyberspace*, pages 62–67, Stanford, March 22-24 1999.

[195] K. Wittenburg, D. Das, W. Hill, and L. Stead. Group asynchronous browsing on the world wide web. In *In Proceeding of the World Wild Web Confernece*, pages 51–62, 1995.

[196] Yahoo. Yahoo directory.

[197] T. Yan and H. García-Molina. Sift : A tool for wide-area information dissemination. In *Proceedings of USENIX Technical Conference*, pages 177–186, New Orleans, Louisiana, January 16-20 1995.

[198] Q. Yang, H. F. Wang, G. Wen, J. R. Zhang, Y. Lu, K. F. Lee, and H. J. Zhang. Toward a next-generation search engine. In *Proceedings of the Sixth Pacific International Conference on Artificial Intelligence*, pages 5–15, Melborne, Australia, August 28 - September 01 2000. Springer.

[199] S. Yu, D. Cai, J. Wen, and W. Ma. Improving pseudo-relevance feedback in web information retrieval using web page segmentation. In *Proceedings of the 12th international conference on World Wide Web*, pages 11–18, Budapest, Hungary, 2003. ACM.

[200] W. N. Zemirli, L. Tamine, and M. Boughanem. Accès personnalisé à l'information : vers la définition d'un profil utilisateur multidimensionnel. In *International Symposium On Programming Systems*, pages 20–28. USTHB, 2005.

[201] W. N. Zemirli, L. Tamine, and M. Boughanem. A personalized retrieval model based on influence diagrams. In *International Conference on Modeling and Using Context - International Workshop on Context Based Information Retrieval*, Roskilde University, Denmark, août 2007. Doan Bich-Liên, Jose Joemon, Melucci Massimo (Eds.), Roskilde University, ISSN 0109-9779.

[202] W. N. Zemirli, L. Tamine, and M. Boughanem. Présentation et évaluation d'un modèle d'accès personnalisé à l'information basé sur les diagrammes d'influence. In *Congrès Informatique des Organisations et Systèmes d'Information et de Décision (INFORSID 2007)*, pages 75–86, mai 2007.

[203] C. Zhai and J. Cohen. Beyond independent relevance : Methods and evaluation metrics for subtopical retrieval. In *Proceedings of the 27th annual international ACM SIGIR Conference on Research and development in Information retrieval*, pages 10–17, August 2003.

Liste des figures

Liste des tableaux