INSTITUT DE RECHERCHE EN INFORMATIQUE DE TOULOUSE

Université Paul Sabatier & Institut National Polytechnique de Toulouse

Formation Doctorale en informatique Année 2003/2004

DEA Informatique de l'Image et du Langage (2IL)
Responsable : René Caubet

Université Paul Sabatier – Toulouse III
118, Rte de Narbonne Laboratoire de l'Institut de Recherche en Informatique de Toulouse
31062 Toulouse Cedex.

Vers le développement d'un système de recherche d'information personnalisé intégrant le profil utilisateur

Par **Zemirli W.Nesrine**

Equipe **SIG/RI**

Directeur de recherche : **Mohand BOUGHANEM**, Professeur
Encadrement et Suivi : **Mohand BOUGHANEM,** Professeur
Linda LECHANI, Maître de conférence

Mots clés : Système de recherche d'information, systèmes adaptatifs, système de recherche d'information personnalisé, profil utilisateur, intégration.

Résumé : Ce rapport traite de la personnalisation de l'accès à l'information pertinente à travers la modélisation et l'intégration d'un profil utilisateur multicritères dans le processus de recherche.
Un état de l'art général sur les différentes approches développées par la communauté de la recherche d'information a permis de faire ressortir les avantages et les inconvénients de chacun des systèmes développés.

Key words: System of search for information, systems adaptive, system of search for information personalized, user profile, integration.

Abstract: This report deals with the personalization of the access to relevant information through the modeling and the integration of a multicriterion user profile in the search process.
A general state of the art on the various approaches developed by the community for information retrieval allowed to emphasize the advantages and the disadvantages of each developed system.

REMERCIEMENTS

Je tiens à remercier très sincèrement Monsieur Claude Chrisment Professeur à l'Université Paul Sabatier et Monsieur Gilles Zurfluh Professeur à l'Université Toulouse I, co-responsables de l'équipe SIG (Systèmes d'Informations Généralisés), pour m'avoir accueillie au sein de leur équipe. Qu'ils trouvent ici l'expression de mon grand respect.

Je tiens à exprimer ma vive reconnaissance et ma très grande considération à M. Mohand BOUGHANEM pour avoir accepté de diriger mes travaux et pour sa disponibilité tout au long de l'année. Son humilité, sa compréhension, son sens des relations, sa rigueur, son expérience et ses critiques constructives, m'ont été précieux.

Ce sentiment va également à l'encontre de mon encadreur, Mme Linda Lechani, qui n'a pas ménagé son temps et ses efforts pour m'orienter, pour ses remarques objectives, ses encouragements ininterrompus, le climat agréable de travail qu'elle crée.

Enfin, je tiens à adresser mes remerciements à tous les membres de l'équipe SIG, pour leur sympathie, leur gentillesse, leur compétence et leurs qualités humaines.

Enfin, je tiens à associer ma famille, à la remercier pour son soutien et de sa précieuse aide tout au long de cette année.

DEDICACES

A la mémoire de mon oncle MOHAMED que je n’oublierai jamais

A ma tendre et douce maman

A mon père

A toute ma famille et mes amis

Je dédie ce travail

TABLE DES MATIERES

Introduction générale……..……..……...1

Chapitre 1: De la recherche d'information à la recherche adaptative

I Introduction ……..……..……...….3

II Principes généraux de la recherche d'information……..4

II.1 Les systèmes de recherche d'informations ……...…… ..…….4

II.1.1. Définition ……...……..…….4

II.1.2 Le processus RI……..…..……5

II.1.2.1 Notions de document et de requête…….……...5

II.1.2.2 Principales phases du processus de RI…….…….....6

II.1.2.2.a L'indexation …….…….6

II.1.2.2.b L'appariement requête-document …..……..…….8

II.2 Les principaux modèles de RI……..…. 9

II.2.1 Le modèle booléen …….…….……9

II.2.2 Le modèle vectoriel …… …… …….10

II.2.3 Le modèle probabiliste ……...……....11

II.3 Evaluation de SRI…….....12

III La recherche d'information adaptative …….……13

III.1 La reformulation de requêtes …..……..……...14

III.1.1 Réinjection de pertinence dans le modèle vectoriel ……...…….. 14

III.1.2 Réinjection de pertinence dans le modèle probabiliste ……..…….. 15

III.2 Le filtrage d'information ……..…..…….. 16

III.2.1 Principe du filtrage …….…… 16

III.2.2 Les modes de filtrage ……17

III.2.2.1 Le filtrage collaboratif…… … …….…….17

III.2.2.2 Le filtrage explicite …….18

III.2.2.3 Le filtrage implicite…...…….……..18

IV Synthèse et conclusion…….…….…….19

Chapitre 2 : Accès personnalisé à l'information

I Introduction…….. …… ……21

II Les systèmes de recherche d'informations personnalisés ……..…….....…….21

II.1 Définition……..... ..…...21

II.2 Notion de profil…….23

II.2.1 Le profil utilisateur ……. …… ……...…23

II.2.2 Le profil document……..…….....23

II.3 Accès personnalisé à l'information……24

II.3.1 Le but de la recherche …….....25

II.3.2 Représentation des résultats de recherche …….…...25

III Mise en œuvre d'un SRIP …….…...26

III.1 Modélisation de l'utilisateur ……....26

III.1.1 Représentation du profil utilisateur……...26

III.1.1.1 Représentation vectorielle …… …….26

III.1.1.2 Représentation hiérarchique ……..…….27

III.1.1.3 Représentation multidimensionnelle ……..28

III.1.2 Construction du profil…….30

III.1.2.1 Analyse statistique des termes ……..…31

III.1.2.2 Techniques d'apprentissage ……...……..31

III.1.2.3 Concept de la vie artificielle …...……..….……32

III.2 Modélisation des documents …….....32

IV Conclusion ….…….……..35

Chapitre 3 : Définition et intégration du profil utilisateur dans le processus de RI pour un accès personnalisé à l'information

I Problématique et objectifs…….36

II Définition du profil utilisateur……..37

II.1 Contenu du profil utilisateur …… ……...38

II.1.1 Catégorie des préférences……...38
II.1.2 Catégorie des données personnelles 40
II.1.3 Catégorie des données de l'environnement ……...41
II.2 Représentation du profil ……..…..42

III Mise en œuvre du SRIP…….……...45

III.1 Le processus d'accès personnalisé……..45
III.1.1 Construction des profils……45
III.1.2 Présélection de l'espace de recherche……45
III.1.3 Evaluation de la requête …….....46
III.1.4 Présentation des résultats……..46
III.1.5 Mise à jour des profils……...46
III.2 L'architecture du SRIP…….....46
III.2.1 Module de construction de profils……47
III.2.1.1 Le gestionnaire du profil utilisateur……..… 47
III.2.1.2 Le gestionnaire du profil document……...48
III.2.2 Module d'intégration de profils ...……..48
III.2.2.1 Le gestionnaire de l'espace de recherche……49
III.2.2.2 Le gestionnaire de la requête……...……...50
III.2.2.3 Le gestionnaire d'appariement……..50
III.2.3 Module de présentation du résultat ……...51
III.2.4 Module d'évolution du profil ……51

IV Interaction du profil utilisateur - processus du SRIP…….52
IV.1 Intégration du profil dans la phase de présélection de l'espace de recherche…...54
IV.2 Intégration du profil dans la phase d'évaluation de requête …….54
IV.3 Intégration du profil dans la phase de présentation du résultat …….54

V Conclusion ……...55

Conclusion générale……..……..…....56

Bibliographie..……..58

Liste des figures

Figure 1.1 Processus en U de recherche d'informations…… ……....5
Figure 1.2 Processus de filtrage d'information.…… …....17
Figure 2.1 Architecture générale d'un SRIP…… ..……22
Figure 3.1 Représentation multidimensionnelle du profil utilisateur…… … ..…..42
Figure 3.2 Architecture générale du SRIP…… …….47
Figure 3.3 Module de construction de profils…… … ……....48
Figure 3.4 Module d'intégration de profil…… …….49
Figure 3.5 Module de représentation du résultat…… ……...51
Figure 3.6 Module d'évolution du profil…… …… ……...51
Figure 3.7 Architecture générale du SRIP…… …… ……....52
Figure 3.8 Correspondance profil utilisateur/module SRIP…… ……...53

INTRODUCTION GENERALE

Le rapide développement des technologies de l'information durant ces dernières années a eu pour conséquence une prolifération de sources d'informations hétérogènes. Il devient de plus en plus difficile pour les utilisateurs de retrouver précisément ce qu'ils recherchent dans cette masse de données. 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.
Le développement d'outils de recherche efficaces, permettant notamment à l'utilisateur de n'avoir accès qu'à l'information qu'il juge pertinente, devient une nécessité absolue.
La première approche suivie pour améliorer les performances des systèmes a été la reformulation de la requête. Elle se base sur le principe que l'utilisateur n'est souvent pas capable de formuler ses besoins en informations et consiste à l'aider dans sa formulation de requête en ajoutant et repondérant des termes de la requête à partir des documents qu'il juge.
L'inconvénient majeur de cette approche est le grand nombre de documents que l'utilisateur doit juger. En outre rien ne garantit que l'utilisateur est prêt à juger tous ces documents ce qui entraîne le plus souvent l'abandon de l'utilisateur et l'échec de la recherche.
L'autre technique développée, est le filtrage d'informations. Elle consiste à concevoir des mécanismes qui permettent de faciliter la tâche de recherche à l'utilisateur, en lui faisant parvenir continuellement l'information qui l'intéresse selon *son profil*
L'inconvénient majeur de ces systèmes est que la saisie manuelle des informations du profil est un processus long et ennuyeux, il y a également une forte surcharge cognitive de l'utilisateur, ce qui entraîne le plus souvent un abandon de sa part.

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 globale de recherche. Les travaux tendent vers des objectifs généraux communs : délivrer l'information pertinente en fonction des caractères spécifiques de l'utilisateur, adapter les résultats de la recherche aux attentes de l'utilisateur et idéalement de les précéder. Ces travaux s'inscrivent dans le cadre précis de la « personnalisation de l'information ».

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.

Répondre aux besoins en information des utilisateurs d'une manière personnelle ne peut se faire sans inclure l'utilisateur dans le processus de RI. Inclure l'utilisateur dans le processus de recherche implique la représentation de ce dernier dans un modèle ou par une structure qui permet son exploitation par le système de recherche d'information.

Cependant, comme le contenu et la structure du profil dépendent fortement de l'application, le problème de capture des paramètres du profil reste encore posé. Le fait que les utilisateurs ont des centres d'intérêts et une demande d'informations différentes et qu'en plus, ils ont souvent des idées floues sur leurs préférences, complique l'extraction des caractéristiques pertinentes et nécessite un processus de découverte de leurs préférences.
Malgré le succès des approches utilisées pour personnaliser les système de recherche, ils présentent certains inconvénients concernant l'initialisation, la représentation, l'exploitation et l'évolution du profil utilisateur. L'utilisateur devrait être représenté par un modèle général regroupant l'ensemble des dimensions informationnelles le caractérisant.

Les problèmes liés à la mise en œuvre d'un système de personnalisation performant concernent la modélisation et l'intégration du profil utilisateur dans le processus de recherche, il faut donc avoir une vision globale de ce qu'est « l'accès personnalisé à l'information » afin de mieux cerner tous les paramètres déterminants de la personnalisation.

La contribution de ce rapport porte sur la proposition d'un modèle de profil multidimensionnel. Ce modèle va permettre de structurer les données nécessaires pour décrire les préférences d'un utilisateur ce qui va nous donner une vision globale sur tous les problèmes liés à la personnalisation de l'information. Ce travail a nécessité un état de l'art approfondi sur l'évolution des performances systèmes de recherche d'information et les approches utilisées pour fournir un accès adaptatif puis personnalisé à l'information.
Ce rapport est organisé en trois chapitres. Le chapitre 1 est consacré à l'état de l'art des approches développées pour améliorer les performances des systèmes de recherche. Il retrace l'évolution des systèmes de recherche d'information classiques vers les systèmes de recherche d'information adaptatifs. Il décrit les avantages et les inconvénients des chacune de ces approches. Le chapitre 2 aborde l'accès personnalisé à l'information, nous y présentons les concepts clés de la personnalisation ainsi que les approches mises en œuvre dans les systèmes de recherche d'informations. La difficulté fût de faire ressortir les différentes approches de personnalisations indépendamment des différents types systèmes existants.
Ensuite, dans le chapitre 3, nous proposons un modèle de profil utilisateur multidimensionnel qui prend en compte tous les aspects de la personnalisation. Nous avons défini les phases d'intégration du profil utilisateur comme composante principale dans le processus de recherche, ainsi qu'une architecture du système de personnalisation. Nous terminons par une conclusion générale et des perspectives de recherche.

Chapitre 1

De la recherche d'information à la recherche adaptative

I Introduction

Le rapide développement des nouvelles technologies de l’information et de la communication ainsi que l’essor du web, nous a confronté à une très grande masse d’informations hétérogènes. Les masses d’informations accessibles n’ont cessé d’augmenter, et les volumes de documents qui les stockent s’accroissent très rapidement. En mars 2002, le plus grand moteur de recherche a contenu approximativement 968 millions de pages classées dans sa base de données [Search 02]. A titre d’exemple, aujourd'hui le moteur de recherche Google inclut 3.083.324.652 sites web [Neuhold 03].

En raison de cette augmentation constante du volume d’informations, nous arrivons à une situation paradoxale : jamais il n’y a eu autant d’informations disponibles, mais trouver dans cette accumulation ce que l’on recherche précisément, devient de plus en plus ardu. Le problème n'est plus tant la disponibilité de l'information mais la capacité de sélection de l’information répondant aux besoins précis d'un utilisateur, à partir des représentations qu'il perçoit.
La conception et la mise en œuvre d’outils efficaces, permettant notamment à l’utilisateur de n’avoir accès qu’à l’information qu’il juge pertinente, devient une nécessité absolue.

Cependant, en dépit des nouvelles technologies qui supportent actuellement les systèmes de recherche d’informations, les utilisateurs ne sont toujours pas satisfaits du processus de recherche et des résultats présentés. Les moteurs de recherche renvoient habituellement plus de 1.500 résultats par question, pourtant parmi les vingt principaux résultats, seulement la moitié d’entre eux sont susceptibles d’être appropriés aux besoins de l'utilisateur.

Les études menées par [Bradford 99] ont montré que la majorité des utilisateurs n’ont aucune idée du fonctionnement des moteurs de recherche qu’ils utilisent, et de ce fait, ils expriment mal leurs besoins. Les utilisateurs n'utilisent généralement que quelques mots (4 ou 5 au maximum) pour formuler leurs requêtes, ce qui donne des spécifications inachevées sur leur besoin en informations.
Du fait de cette inexpérience, les performances des processus de recherche sont amoindries. L’utilisateur est de plus en plus insatisfait et un sentiment de frustration naquit vis à vis du système de recherche.

Un autre facteur de l’insatisfaction des utilisateurs, mis à part leur inexpérience et le manque de connaissances de leur besoin informationnel effectif est, que la majorité des systèmes de recherche disposent de peu d’informations sur les utilisateurs pouvant améliorer le processus de recherche.

Il n'y a généralement pas de mécanisme explicite qui permet d'inclure l'utilisateur dans le processus de recherche. Il n'y a également pas d'informations à priori qui indiquent quel ensemble de documents est le plus probable d'inclure l'information pertinente pour l'utilisateur. La pertinence étant une notion complètement subjective et dépendante de l'utilisateur.

De ce constat sont nées de nouvelles approches pour la conception et la mise en œuvre des systèmes de recherche d'informations. Dans le but d'améliorer la satisfaction de l'utilisateur, il est opportun de lui offrir un **accès personnalisé** à l'information appropriée et de l'**assister** lors de sa recherche. Ce n'est plus la **quantité** mais la **qualité** des informations fournies qui détermine si un utilisateur est satisfait par un système ou pas.

La personnalisation des systèmes de recherche d'informations nécessite ainsi l'élaboration de modèles et de mécanismes représentant et incluant l'utilisateur dans le processus de recherche.
Les principales questions qui se posent lors de la conception d'un système de recherche d'informations personnalisé sont alors :

- Quelles sont les propriétés qui caractérisent un profil d'utilisateur ?
- Comment modéliser puis faire évoluer ce profil ?
- Comment intégrer le profil utilisateur dans le processus de recherche ?

Ce chapitre tente de répondre à ces questions. A cet effet, on présente, dans la première section, les concepts généraux, mais fondamentaux, de la recherche d'informations classiques. La deuxième section aborde les différentes approches proposées pour adapter le processus de recherche aux utilisateurs. La dernière section présente une synthèse de ces différentes approches ainsi que les limites de chacune.

II Principes généraux de la recherche d'informations

La recherche d'informations (RI) fournit des techniques et des outils pour trouver les documents contenant l'information pertinente aux besoins des utilisateurs.

II.1 Les systèmes de recherche d'informations

II.1.1 Définition

*Un **Système de Recherche d'Informations (SRI)** est un système informatique qui permet de retourner à partir d'un ensemble de **documents**, ceux dont le contenu **correspond** le mieux à un **besoin** en informations d'un utilisateur, exprimé à l'aide d'une requête.*

Un SRI inclut un ensemble de procédures et d'opérations qui permettent la gestion, le stockage, l'interrogation, la recherche, la sélection et la représentation de cette masse d'informations.

II.1.2 Le processus de RI

Les différentes étapes du processus de RI, sont représentées schématiquement par le processus en U (voir *Fig. I.1*) [Belkin 92]. La figure illustre particulièrement :

- les notions de documents et de requêtes qui sont des conteneurs d'informations,
- les opérations d'analyse, d'indexation et d'appariement qui permettent globalement de traiter la requête dans le but de sélectionner des documents à présenter à l'utilisateur.

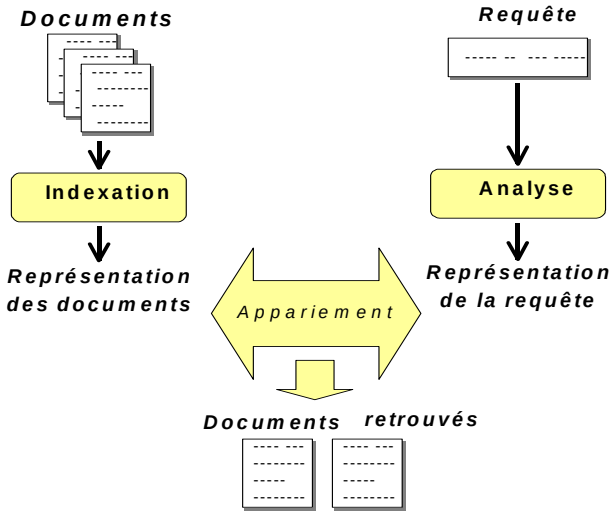

Fig 1.1. *Processus en U de recherche d'informations*

II.1.2.1 Notions de document et de requête

➢ **Document**

Le document représente le conteneur élémentaire d'information, exploitable et accessible par le SRI. Un document peut être un texte, une page WEB, une image, une bande vidéo, etc. Dans notre contexte, nous appelons document toute unité qui peut constituer une réponse à un besoin en information exprimé par un utilisateur.

➢ **Requête**

Une requête constitue l'**expression** du **besoin** en informations de l'utilisateur. Plusieurs systèmes utilisent des langages différents pour décrire la requête :

- par une liste de mots clés : cas des systèmes SMART [Salton 71] et Okapi [Robertson 9],
- en langage naturel : cas des systèmes SMART [Salton, 71] et SPIRIT [Fluhr 85],
- en langage booléen : cas du système DIALOG [Bourne 79],
- en langage graphique : cas du système NEURODOC [Lelu 92].

II.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. Ceci induit deux principales phases dans le déroulement du processus : indexation et appariement requête/documents.

II.1.2.2.a L'indexation

Un SRI gère les différentes collections de documents en les organisant sous forme d'une représentation intermédiaire permettant de refléter aussi fidèlement que possible leur contenu sémantique. L'interrogation de ce fond documentaire à l'aide d'une requête nécessite également la représentation de cette dernière sous une forme compatible avec celle des documents. Ce processus de conversion est appelé **indexation** (également appelé analyse pour la requête).

L'indexation est une étape très importante dans le processus de RI. Elle consiste à déterminer et extraire les termes représentatifs du contenu d'un document ou d'une requête, qui couvrent au mieux leur contenu sémantique. 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 requête. Ce dernier est souvent une liste de termes ou groupe de termes significatifs pour l'unité textuelle correspondante, généralement assortis de poids représentant leurs degré de représentativité du contenu sémantique de l'unité qu'ils décrivent.
Les descripteurs des documents (mots, groupe de mots) sont rangés dans une structure appelée dictionnaire constituant le **langage d'indexation**.

Un groupe de mots est à priori sémantiquement plus riche que les mots qui le composent pris séparément. Cet argument conduit à ne pas considérer simplement les mots simples comme unités de base dans le langage d'indexation mais également des groupes de mots. Ce groupe de mots forme ce que l'on appelle un *thesaurus*. Ce dernier inclut des relations de type linguistiques (équivalence, association, hiérarchisation) et statistiques (pondération) [Mothe 00].

L'indexation peut être caractérisée par son mode et fonction de pondération

- **Mode d'indexation**

L'indexation peut être manuelle, automatique ou semi-automatique :
- *manuelle* : chaque document est analysé par un spécialiste du domaine correspondant ou par un documentaliste,
- *automatique* : chaque document est analysé à l'aide d'un processus entièrement automatisé,

- *semi-automatique* : le choix final reste au spécialiste du domaine correspondant ou documentaliste, qui intervient souvent pour établir des relations sémantiques entre mots-clés et choisir les termes significatifs.

❑ **Fonction de pondération**

La pondération permet d'affecter à chaque terme d'indexation une valeur qui mesure son importance dans le document où il apparaît.
Le pouvoir de discrimination des termes pour décrire le contenu des documents n'est pas identique pour tous les termes. Pour trouver les termes du document qui représentent le mieux son contenu sémantique, [Roberston 76] 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 [Robertson 76], [Singhal 97] et [Sparck Jones 79].

On y distingue :

- *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. $0.5 + 0.5 \quad \frac{Tf}{Max(Tf)}$

- *Idf* (*Inverse of 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. $Idf = \log\left(\frac{N}{df}\right)$,
2. $Idf = \log\left(\frac{N - df}{df}\right)$.

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:

$$Tf.Idf = \left(0.5 + 0.5\frac{Tf}{Max(Tf)}\right) * \log\left(\frac{N}{df}\right)$$

Une normalisation de la mesure du *Tf.Idf* par rapport à la longueur des documents a été proposée par [Singhal 95].

$$Tf.Idf = \frac{Tf + \log\left(\frac{N - df + 0.5}{df + 0.5}\right)}{2.\left(0.25 + 0.75.\frac{dl}{\Delta d}\right)}$$

Idf est la longueur du document en nombre de termes et Δd la longueur moyenne des documents de la collection.
En effet, lors des compagnes d'évaluation internationales, la mesure a eu des performances très limitées dans des corpus de taille très variable. Le problème posé est que les termes appartenant aux documents longs apparaissent très fréquemment et emportent le poids sur les termes appartenant à des documents moins longs. Les documents longs auront alors plus de chance d'être sélectionnés [DeClaris 94].

II.1.2.2.b L'appariement requête-document

Les SRI intègrent un processus de recherche/décision qui permet de sélectionner l'information jugée pertinente pour l'utilisateur. A cet effet, une mesure de similitude (correspondance) entre la requête indexée et les descripteurs des documents de la collection est calculée. Seuls les documents dont la similitude dépasse un seuil prédéfini sont sélectionnés par le SRI.
La fonction de correspondance est un élément clé d'un SRI, car la qualité des résultats dépend de l'aptitude du système à calculer une pertinence des documents la plus proche possible du jugement de pertinence de l'utilisateur [Tmar 02].

Il existe deux types d'appariement :

- *Appariement exact*

Le résultat est une liste de documents respectant exactement la requête spécifiée avec des critères précis. Les documents retournés ne sont pas triés.

- *Appariement approché*

Le résultat est une liste de documents sensés être pertinents pour la requête. Les documents retournés sont triés selon un ordre de mesure. Cet ordre reflète le degré de pertinence document/requête

II.2 Les principaux modèles de RI

Un modèle de recherche d'informations est formellement décrit par un quadruplet **[D, Q, F, R (q_i ,d_j)]** [Yates 99] où :

D : ensemble des représentants des documents de la collection,

Q : ensemble des représentants des besoins en informations,

F : schéma du support de représentation des documents, requêtes et relations associées.

R (q_i ,d_j) : fonction d'ordre associée à la pertinence.

Nous présentons ici les modèles les plus couramment utilisés pour la RI, notamment le modèle booléen, le modèle vectoriel et le modèle probabiliste. [Grossman 98] détaille les différents modèles de RI.

II.2.1 Le modèle booléen

C'est le premier modèle utilisé en RI [Salton 71] ; il est basé sur la théorie des ensembles et l'algèbre de Boole [Gessler 93]. Le modèle booléen propose la représentation d'une requête sous forme d'une équation logique. Les termes d'indexation sont reliés par des connecteurs logiques *ET*, *OU* et *NON*.

L'approche booléenne consiste à trouver les documents qui ont *exactement* les mêmes termes qu'une requête construite par mots clefs. Les requêtes peuvent être affinées grâce aux opérateurs *OR* ou *AND* ou encore au moyen d'opérateurs comme *NEAR*. Ce type de recherche est à la base des moteurs de recherche comme Altavista[1] ou Google[2].

Cette approche est très efficace pour des requêtes utilisant des termes très spécifiques ou portant sur des domaines techniques particuliers avec leur vocabulaire propre mais son intérêt reste néanmoins limité.

L'inconvénient majeur du modèle booléen réside dans sa caractéristique de fournir une réponse binaire (les documents contiennent les termes demandés ou ne les contiennent pas). Ceci induit un volume de réponses important sans ordre spécifique des documents résultants.

---

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

[2] http://www.google.com

Les deux modèles présentés ci-dessous largement utilisés en pratique, permettent de remédier à ces inconvénients.

II.2.2 Le modèle vectoriel

Le modèle vectoriel introduit par Salton [Salton 71], repose sur les bases mathématiques des espaces vectoriels. Dans ce modèle, les documents et les requêtes sont représentés dans un espace vectoriel engendré par l'ensemble des termes d'indexation $t_1, t_2, ..., t_T$ Où N est le nombre total de termes issus de l'indexation de la collection des documents.

Chaque document est représenté par un vecteur : $D_j = (d_{1j}, d_{2j}, ...d_{ij} ....., d_{Tj})$
Chaque requête est représentée par un vecteur : $Q = (q_1, q_2, ...q_i ..., q_T)$

Avec :
d_{ij} : Poids du terme t_i dans le document D_j ,
q_i : Poids du terme t_i dans la requête Q.

Les termes de poids nul représentent les termes absents dans un document alors que les poids positifs représentent les termes assignés.
La fonction de calcul du coefficient de similarité entre chaque document Dj, représenté par le vecteur $(d_{1j}, d_{2j}, ..., d_{Tj})$, et la requête Q, représentée par le vecteur $(q_1, q_2, ..., q_T)$ est appelée *Retrieval Status Value* ou *RSV*.
Ce coefficient de similarité est calculé sur la base d'une fonction qui mesure la colinéarité des vecteurs document et requête. On peut citer notamment les fonctions suivantes :

- **Produit scalaire :** $$RSV(Q, D_j) = \sum_{i=1}^{T} q_i * d_{ij}$$

- **Mesure de Jaccard :** $$RSV(Q, D_j) = \frac{\sum_{i=1}^{T} q_i * d_{ij}}{\sum_{i=1}^{T} q_i^2 + \sum_{i=1}^{T} d_{ij}^2 - \sum_{i=1}^{T} q_i * d_{ij}}$$

- **Mesure de cosinus :** $$RSV(Q, D_j) = \frac{\sum_{i=1}^{T} q_i * d_{ij}}{\left(\sum_{i=1}^{T} q_i^2\right)^{1/2} * \left(\sum_{i=1}^{T} d_{ij}^2\right)^{1/2}}$$

La similarité entre deux textes (requêtes ou documents) dépend ainsi des poids des termes coïncidant dans les deux textes. Il est donc possible de classer les documents par ordre de pertinence décroissante

L'avantage du modèle vectoriel par rapport au modèle booléen réside particulièrement dans l'ordonnancement des documents sélectionnés selon leurs pertinences. Cependant, l'inconvénient majeur de l'approche vectorielle réside dans le fait que l'association entres les termes d'indexation n'est pas considérée. Il est impossible de représenter des phrases ou des mots multi termes. On considère effectivement que les termes sont indépendants [Yates 99].

II.2.3 Le modèle probabiliste

Le modèle probabiliste a été proposé par Robertson et Sparck Jones [Robertson 76], il utilise un modèle mathématique fondé sur la théorie de la probabilité conditionnelle (appelé aussi modèle de la théorie de pertinence). Lors du processus d'indexation deux probabilités conditionnelles sont utilisées :

$P\left(t/_{Pert}\right)$: La probabilité pour que le terme *t* apparaisse dans un document donné sachant que ce document est pertinent pour la requête.

$P\left(t/_{NonPert}\right)$: La probabilité pour que le terme *t* apparaisse dans un document donné sachant que ce document est non pertinent pour la requête.

En supposant que la distribution des termes dans les documents pertinents est la même que leur distribution par rapport à la totalité des documents, et que les variables "document pertinent" et "document non pertinent" sont indépendantes, la fonction de recherche est obtenue en calculant la probabilité de pertinence d'un document $P(\text{Pert} / \text{D})$

$$P(Pert / D) = \frac{P(D / Pert) * P(Pert)}{P(D)},$$

$$P(NonPert / D) = \frac{P(D / NonPert) * P(NonPert)}{P(D)},$$

$$\text{Avec } P(D) = P(D / Pert) * P(Pert) + P(D / NonPert) * P(NonPert),$$

Où :

P(D/Pert) (respectivement *P(D/NonPert)*) : Probabilité d'observer D sachant qu'il est pertinent (respectivement non pertinent).

P(Pert) (respectivement *P(NonPert)*) : Probabilité à priori qu'un document soit pertinent (respectivement non pertinent).
Le coefficient de similarité requête document (*RSV*) peut être calculé par différentes formules. Robertson et Spark-Jones [Robertson 96] proposent la formule suivante :

$$RSV = \sum \log\left(\frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)} \right)$$

Où
N: nombre total de documents de la base,
n: nombre de documents contenant le terme,
R: nombre de documents connus comme étant pertinents,
r: nombre de documents connus comme étant pertinents et contenants le terme.
L'ajout de 0.5 à tous les membres s'explique par la nécessité d'écarter tous les cas limites qui entraîneraient des valeurs nulles de ces membres.

Ce modèle a donné lieu à de nombreuses extensions. Il est à l'origine du système OKAPI qui est l'un des systèmes les plus performants selon les compagnes d'évaluation TREC[3] [Walker 97].
L'inconvénient majeur de ce modèle est que les calculs des probabilités sont complexes et que l'indépendance des variables n'est pas toujours vérifiée voir pas prise en compte.

II.3 Evaluation de SRI

La démarche de validation en RI se base sur l'évaluation expérimentale des performances du modèle ou du système proposé. L'évaluation des performances d'un modèle de RI, permet de paramétrer le modèle, d'estimer l'impact de chacune de ses caractéristiques et de fournir des éléments de comparaison entre modèles.

Cette évaluation peut porter sur plusieurs critères : le temps de réponse, la pertinence, la qualité et la présentation des résultats, etc. Le critère le plus important est celui qui mesure la capacité du système à satisfaire le besoin en information de l'utilisateur, c'est à dire la pertinence. Deux facteurs permettent d'évaluer ce critère. Le premier est le **taux de rappel**, il mesure la capacité du système à sélectionner tous les documents pertinents. Le second est le **taux de précision**, il mesure la capacité du système à rejeter tous les documents non pertinents. On calcule :

$$Taux\_rappel = R_d = \frac{\text{Nombre de documents pertinents restitués}}{\text{Nombre de documents pertinents}}$$

[3] WWW.TREC.com

$$Taux\_précision = P_d = \frac{\text{Nombre de documents pertinents restitués}}{\text{Nombre de documents restitués}}$$

Ce type de mesure est effectué sur des collections de tests. Une collection de tests est composée d'un ensemble de documents, d'un ensemble de requêtes et d'un ensemble de jugements de pertinence.

L'initiative la plus importante actuellement pour la construction de collections de tests est sans conteste TREC (Text REtrieval Conference, http://trec.nist.gov) [Harman 92]. TREC est plus qu'une collection de tests, c'est un programme d'évaluation des SRI, initié par le NIST (National Institute of Standards and Technology) aux USA. TREC fournit une plate-forme comportant des collections de tests, des tâches spécifiques et des protocoles d'évaluation pour chaque tâche, pour l'évaluation et la comparaison d'expérimentations sur des collections volumineuses de textes. Il faut noter que les collections TREC représentent aujourd'hui un référentiel incontournable en RI [TREC-9 00].

III La recherche d'informations adaptative

L'efficacité et la qualité des mécanismes (indexation et appariement requête/documents) mis en œuvre durant le processus de recherche ont un impact direct et déterminant sur les performances d'un SRI, en particulier sur la qualité des réponses.

Lors de l'appariement requête/documents, 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, les performances des SRI ne dépendent pas seulement des termes d'indexation et du mécanisme d'appariement mais aussi de façon non négligeable de l'*utilisateur*.

En effet, quand l'utilisateur formule son besoin en information à l'aide d'une requête, les termes choisis dans cette requête ont une grande influence sur la réponse du système. Lorsque l'utilisateur initie la recherche, c'est dans le but de combler un manque d'informations sur un sujet précis. Ce besoin d'informations est fortement lié à une activité ponctuelle de l'utilisateur, au contexte et à l'environnement où il effectue sa recherche. Mais il est, également, lié aux préférences et centre d'intérêts de l'utilisateur. En fin de compte pour qu'un SRI soit performant, il faudrait qu'il arrive à satisfaire l'utilisateur qui est le seul juge de la qualité des résultats fournis.

En considérant ces paramètres de performance, de nombreux travaux ont visé la conception de SRI dits *adaptatifs*, en ce sens qu'ils exploitent des informations extraites de l'utilisateur dans le but de s'y adapter. Ces systèmes adaptent généralement une ou plusieurs phases du processus de recherche d'informations à l'utilisateur.

Pour notre part, on s'intéresse particulièrement à l'accès adaptatif à l'information. Dans ce contexte, on présentera un aperçu des techniques de reformulation de requêtes, et du filtrage d'informations.

III.1 La reformulation de requêtes

La reformulation de requêtes est l'une des méthodes élaborées pour l'adaptation des SRI aux besoins des utilisateurs. C'est un processus ayant pour objectif de générer une nouvelle requête plus adéquate que celle initialement formulée par l'utilisateur.

L'une des stratégies de reformulation de requêtes est celle qui est dirigée par l'utilisateur. Le principe de cette stratégie est de construire une nouvelle requête à partir de la structure des documents jugés par l'utilisateur : c'est ce que l'on appelle la réinjection de pertinence « relevance feedback » [Rochio 71] [Harman 92] [Boughanem 98].

C'est un processus évolutif et interactif. Son principe fondamental est d'utiliser la requête initiale pour amorcer la recherche, puis modifier celle-ci à partir des jugements de pertinence et/ou de non-pertinence de l'utilisateur dans le but de repondérer les termes de la requête initiale, ou y ajouter (respectivement supprimer) d'autres termes contenus dans les documents pertinents (respectivement non pertinents). La nouvelle requête obtenue à chaque itération de feedback, permet de corriger la direction de la recherche dans le sens des documents pertinents.
En effet, la simple comparaison du contenu de la requête et des documents de la base ne permet pas d'avoir tous les documents correspondant à une requête donnée. Il reste toujours des documents pertinents non restitués, car ne contenant pas les termes de la requête.

La reformulation de requête par injection de pertinence a été introduite dans différents modèles de recherche.

III.1.1 Réinjection de pertinence dans le modèle vectoriel

Dans le modèle vectoriel, la requête et les documents sont représentés sous forme vectorielle, 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_{t+1} est construite grâce à la **formule de Rocchio** [Rocchio 71] 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 :

$$Q_{t+1} = \alpha . Q_t + \frac{\beta}{n} \sum_{n_p} D_p^{(t)} - \frac{\gamma}{n_{np}} \sum_{n_{np}} D_{np}^{(t)}$$

Avec :

Q_t : vecteur de la requête initiale,

Q_{t+1}: vecteur de la nouvelle requête,

$D_p^{(t)}$ (resp. $D_{np}^{(t)}$): vecteur d'un document pertinent (resp. non pertinent),

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

α, β, γ : constantes.

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*).

III.1.2 Réinjection de pertinence dans le modèle probabiliste

Sur la base du modèle probabiliste, Robertson et Sparck-Jones [Robertson 76] 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 :

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

Avec

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.

Croft [Croft 83] a défini une méthodologie de repondération en utilisant une version révisée de la formule de pondération de Sparck-Jones :

- Recherche initiale : $w_{ijk} = (C + idf_i).f_{ik}$
- Feedback : $w_{ijk} = \left[C + \log \frac{p_{ij}(1 - q_{ij})}{q_{ij}(1 - p_{ij})} \right].f_{ik}$

Avec

w_{ijk} : Le poids du termes t_i dans la requête Q_j et du document D_k,

idf_i : Fréquence absolue du terme t_i dans la collection,

p_{ij}

: Probabilité que le terme t_i soit assigné à un ensemble de documents pertinents pour une requête Q_j ;

$p_{ij} = \frac{r + 0{,}5}{R + 0{,}1}$ si r>0 ; $p_{ij} = 0{,}01$ si r=0,

q_{ij} : Probabilité que le terme t_i apparaisse dans un ensemble de documents non pertinents pour une requête Q_j ;

$q_{ij} = \frac{n - r + 0{,}5}{N - R + 1{,}0}$ $f_{ik} = K + (1 - K).\frac{freq_{ik}}{\max(freq_k)}$

Où $freq_{ik}$: La fréquence du terme dans le document *k*,

$\max(freq_k)$: La fréquence maximale d'un terme t_i dans le document

C, K : Constantes.

III.2 Le filtrage d'information

Une des approches pour adapter le processus de recherche à l'utilisateur est de *filtrer* l'information en fonction de son profil.

III.2.1 Principe du filtrage

Le filtrage d'informations consiste à concevoir des mécanismes destinés à faire parvenir à l'utilisateur l'information qui l'intéresse directement. Habituellement, on considère qu'un SRI a pour fonction « d'amener à l'utilisateur les documents qui vont lui permettre de satisfaire ses besoins en information » [Belkin 92].

Un système de filtrage d'informations (SFI) peut être assimilé à un assistant personnel et doit être capable d'identifier les documents qui correspondent ou pas aux besoins en information des utilisateurs [Boughanem 04].

Un SFI extrait et achemine au sein d'un grand nombre d'informations, provenant de sources dynamiques, les seuls documents susceptibles de répondre aux besoins et intérêts de l'utilisateur après que celui-ci ait dressé un profil de requête, c'est-à-dire défini ses centres d'intérêts [Amati 97] [Allen 90] [Croft 93].

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 les maintenir continuellement informés des nouveautés relatives à leurs domaines de spécialisation [Luhn 57].

A partir de 1982, Denning [Denning 82] 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 [Foltz 90].

La figure 1.2 schématise le processus de filtrage d'informations. Il débute avec des individus ou groupes d'individus qui ont des intérêts relativement stables à long terme et décrits à travers des profils. La source d'informations provient des producteurs de textes (exemple : journaux, Internet, CD-ROM, etc.). Ces derniers doivent distribuer ces informations aux personnes intéressées. Cette opération est réalisée en comparant les textes aux profils des différents individus.

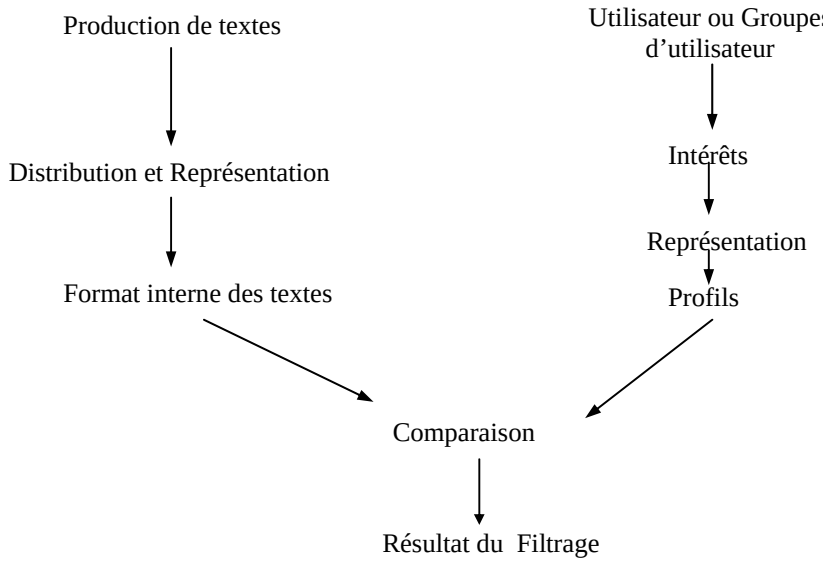

Fig.1.2. *Processus du filtrage d'information*

III.2.2 Les modes de filtrage

On distingue principalement trois modes de filtrage : le filtrage collaboratif, le filtrage explicite et le filtrage implicite.

III.2.2.1 Le filtrage collaboratif

On s'oriente actuellement vers le modèle dit de filtrage par collaboration *(collaborative filtering*). Le terme de filtrage collaboratif décrit les techniques d'un groupe d'utilisateurs pour prédire la préférence inconnue d'un nouvel utilisateur ; les recommandations pour le nouvel utilisateur sont basées sur ces prédictions [Goldberg 00].

Les utilisateurs d'un système d'informations 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 consultés. L'utilisateur donne son avis sur les documents lus et ces réactions peuvent être annotées et consultées par d'autres. Ainsi on établit des relations document-document et document-utilisateur. Le filtre dispose donc d'informations variées sur un document et peut par exemple proposer à un utilisateur la liste des personnes travaillant sur le même sujet et lui sélectionner les documents qu'elles ont consultés. Si parmi ces documents certains sont pertinents, le filtre pourra trouver l'ensemble des documents ayant les mêmes annotations [Miller 97] [Poo 03].

III.2.2.2 Le filtrage explicite

Dans le cas du filtrage *explicite*, le contenu du profil est ciblé en fonction des informations données volontairement par l'utilisateur lors de son enregistrement sur le site. Par exemple, de nombreux sites demandent à l'utilisateur de préciser ses centres d'intérêts. Il recevra ensuite des offres qui lui correspondent.

Les portails tels que web MyYahoo [YHO 04], InfoQuest [Inf 04], FireFly [Fir 04], Bloomberg [Blo04], News Channel [Cnn04] ou PointCast [Poi 04] fournissent des liens vers différentes sources d'informations. Ils permettent la spécification des centres d'intérêts d'un utilisateur par une liste de mots clés. Ils proposent également à l'utilisateur un menu où il peut choisir le contenu de la page en cochant les catégories jugées intéressantes parmi une liste donnée (actualités, horoscopes, programmes télé, météo, cinéma, etc.). Une fois les catégories intéressantes choisies, le système offre à l'utilisateur une page personnelle qui contient des informations dont les thèmes sont les catégories et des liens vers des données supplémentaires.

Une autre technique de filtrage explicite est l'utilisation du feedback de l'utilisateur. En fonction des pages et documents jugés pertinents par l'utilisateur, le système va extraire une liste de termes qui représente le mieux la sémantique des documents. En fonction des données recueillies, le système va pouvoir proposer à l'utilisateur les pages ou des liens vers des pages qui concernent son profil. Les systèmes Syskill et Webert [Pazzani 96] se classent dans ce type.
L'aide à la navigation adaptative du Wisconsin (WAWA) [Shavlik 98] emploie également le feedback utilisateur explicite pour former les réseaux neurologiques afin d'aider les utilisateurs pendant la lecture rapide.

III.2.2.3 Le filtrage implicite

Dans le cas du filtrage *implicite*, le contenu du profil est ciblé en fonction des données extraites automatiquement à partir du comportement de l'utilisateur. Par exemple : une page ouverte pendant un long moment peut signifier l'intérêt de l'utilisateur pour les informations qui y sont présentes. On prend ainsi en compte toutes sortes d'opérations, réalisées à l'aide de la souris, qui vont permettre d'en savoir un peu plus sur l'utilisateur : sélection du texte, passage sur un lien, clic sur un lien, vitesse de défilement de la page, etc.

L'ensemble du contenu d'un site est découpé en segments d'informations que l'on va proposer à l'utilisateur selon les informations recueillies à son sujet. Il faut faire attention à ne pas effectuer une segmentation trop fine, pour garantir un ajustement valable du contenu proposé.

La technique de l'enregistrement Web (Web recording) participe également au filtrage implicite. Elle consiste à mémoriser l'opération de navigation effectuée par un ou plusieurs clients, afin de la reproduire automatiquement pour un autre client. Puisque l'opération est connue, il devient possible de la reproduire à plus grande vitesse. On anticipe sur ce que va faire l'utilisateur pour lui faire gagner du temps.

IV Synthèse et conclusion

La constatation que l'on peut faire actuellement, est que l'on a de plus en plus recours aux moteurs de recherche pour accéder à l'information. Le rapide développement et la maturité de ces systèmes, durant ces dernières années, nous amène cependant à être plus exigent en matière de performances du processus de recherche et de la qualité de l'information retournée.
Face à ce besoin accru de satisfaction, plusieurs approches furent développées, par la communauté de RI, afin de mieux répondre aux attentes de ces utilisateurs et améliorer les performances des SRI.
Les premières approches, s'inscrivant dans le cadre de la RI, sont la reformulation de requêtes et le filtrage d'informations.

La **reformulation de requête** par injection de pertinence est la première technique développée pour s'adapter aux besoins de l'utilisateur. Elle consiste à aider l'utilisateur dans sa formulation de requête en ajoutant des termes à la requête à partir des documents qu'il juge.
La reformulation de requête par injection de pertinence permet effectivement d'améliorer les performances de la recherche [Buckley90] [Salton 89].

Cependant, elle entraîne la surcharge cognitive de l'utilisateur. Ceci est dû au grand nombre de documents que doit juger l'utilisateur. En outre, rien ne garantit que l'utilisateur soit prêt à passer du temps pour juger tous les documents retournés [Vassiliou 02]. Ceci entraîne parfois l'abandon de l'utilisateur et l'échec de la recherche.

L'autre technique développée, est le **filtrage d'informations.** Elle consiste à concevoir des mécanismes qui permettent de faciliter la tâche de recherche à l'utilisateur, en lui faisant parvenir continuellement l'information qui l'intéresse selon *son profil.*

Dans le cas du *filtrage collaboratif,* on se base sur l'hypothèse que les utilisateurs des SRI devraient pouvoir se servir de ce que d'autres ont déjà trouvé et évalué. Il devient possible de traiter n'importe quelle forme de contenu et de diffuser des ressources non nécessairement similaires à celles déjà reçues.

Les systèmes de filtrage collaboratifs organisent les informations des utilisateurs en des groupes d'intérêt semblables, pour prédire les préférences inconnues d'un nouvel utilisateur et de ce fait permettent la recommandation des documents considérés intéressants par d'autres membres de ce groupe.
Cependant leur efficacité dépend fortement du degré de corrélation des utilisateurs du groupe [Poo 03]. De plus, les problèmes subsistent pour les nouveaux documents ; ils ne peuvent être diffusés que si un minimum d'informations les concernant est collecté à partir de l'avis de l'un des utilisateurs. D'un autre côté, les personnes ayant des préférences peu fréquentes risquent de ne pas recevoir de propositions. Ces deux problèmes sont en réalité liés à la taille et à la composition de la population d'utilisateurs.

Dans le cas du *filtrage explicite,* l'information représentant les centres d'intérêts de l'utilisateur sont extraites de manière manuelle. C'est une approche fortement utilisée par les portails du web. Elle permet d'amorcer le processus filtrage. Après que l'utilisateur se soit enregistré et décrit son profil, le système est dans la mesure de lui présenter des documents qui correspondent à ses préférences.
L'inconvénient majeur de ces systèmes est que la saisie manuelle des informations est un processus long et ennuyeux, il y a également une forte surcharge cognitive de l'utilisateur, ce qui entraîne le plus souvent un abandon de sa part.
Une conséquence directe est que, le système obtient très peu d'informations sur l'utilisateur, ce qui limite la recherche de documents pertinents pour ce dernier.
De plus, ces systèmes tendent à être plutôt statiques. Lorsque les centres d'intérêts de l'utilisateur changent et évoluent dans le temps, le système n'a aucun mécanisme automatique pour détecter ces changements, sauf si l'utilisateur effectue lui même la mise à jour de son profil et ce, de manière manuelle. Par conséquent, les performances du système se dégradent avec le temps et les performances d'adaptation du processus de filtrage aux besoins de l'utilisateur sont peu effectives.

Quant aux systèmes de *filtrage implicite*, ils présentent le double avantage de s'adapter à l'utilisateur sans nécessiter une participation active de sa part.
L'utilisateur n'est plus contraint de remplir les formulaires afin de construire son profil, mais c'est le système qui, à l'aide d'agents intelligents, va apprendre et mettre à jour le profil de l'utilisateur en se basant sur ses différents comportements lors de sa recherche et navigation.

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. Les travaux tendent vers des objectifs généraux communs : **délivrer l'information pertinente en fonction des caractères spécifiques de l'utilisateur, adapter les résultats de recherche aux attentes de l'utilisateur et idéalement de les précéder.** Ces travaux s'inscrivent dans le cadre précis de la « personnalisation de l'information ».

Chapitre 2

Accès personnalisé à l'information

I Introduction

Les SRI adaptatifs existants offrent des avantages indéniables pour l'accès à l'information de manière spécifique et adaptative à chaque utilisateur. Ces systèmes constituent une avancée concrète vers l'amélioration des services d'accès à l'information.
Cependant en dépit de ces avancées récentes, les SRI adaptatifs actuels se trouvent de plus en plus confrontés aux exigences sans cesse croissantes des utilisateurs.
En effet, le processus psychologique de satisfaction des êtres humains est tellement complexe et dépend de facteurs peu paramétrables, qu'il est difficile pour un processus informatique, tel que celui mis en œuvre dans les SRI adaptatifs, de fournir l'information pertinente qui satisfait intégralement l'utilisateur.
Rappelons à cet effet que la pertinence de l'information n'est pas une mesure objective, généralisable à tous les utilisateurs. Elle se définit par un ensemble de critères et de préférences **personnalisables** spécifiques à chaque utilisateur ou communauté d'utilisateurs et dépend notamment du centre d'intérêts, du moment et du lieu que l'utilisateur a choisi pour accéder à l'information. La mesure de la pertinence d'une information peut se faire sur le contenu des données (leur sémantique), la qualité du contenant ou du processus de production de contenus.

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. La personnalisation facilite et assiste l'utilisateur lors de sa recherche d'informations [Perenon 00].

Ce chapitre est décomposé en deux grandes parties. La première partie est consacrée aux concepts clés de la personnalisation et la seconde aux approches de mise en œuvre des systèmes de recherche d'informations.

II Les systèmes de recherche d'information personnalisés

Nous présentons dans ce qui suit les principales composants d'un SRI personnalisé.

II.1 Définition

Un système de recherche d'informations personnalisé (SRIP) est un SRI qui intègre totalement l'utilisateur tout au long du processus de recherche. Il répond ainsi de manière personnelle aux besoins en informations de chaque utilisateur.
La RI personnalisée est une activité faisant intervenir deux entités : les caractéristiques de l'utilisateur communément appelés «profil de l'utilisateur » et le caractéristiques des documents appelés les « métadonnées des documents ».
Un SRI personnalisé inclut :

- Des modèles et algorithmes pour capturer et modéliser le but, les préférences et les centres d'intérêts de l'utilisateur ou un groupe d'utilisateur. Un modèle de profil est alors décrit et instancié.

- Une procédure de mise à jour du profil qui traduit son évolution dans le temps.
- Des mécanismes pour extraire les caractéristiques descriptives des documents à l'aide de métadonnées.

Le processus de personnalisation s'effectue en incluant le profil utilisateur et le profil document dans l'ensemble de étapes du processus en U de la RI (figure 1.3). Cette personnalisation peut avoir lieu à différents niveaux du processus :

- Lors de **l'analyse de la requête**. Le SRIP intègre le profil utilisateur pour mieux cibler le besoin informationnel effectif de l'utilisateur.
- Lors de **l'indexation du corpus** documentaire. Le SRIP utilise des métadonnées des documents pour une meilleure représentation de la sémantique des documents.
- Lors de **l'appariement requête/document**. Le SRIP inclut également le profil utilisateur pour calculer la pertinence d'un document.
- Lors de **l'affichage des résultats**. Le SRIP restitue les documents selon la directive et les préférences incluses dans le profil utilisateur.

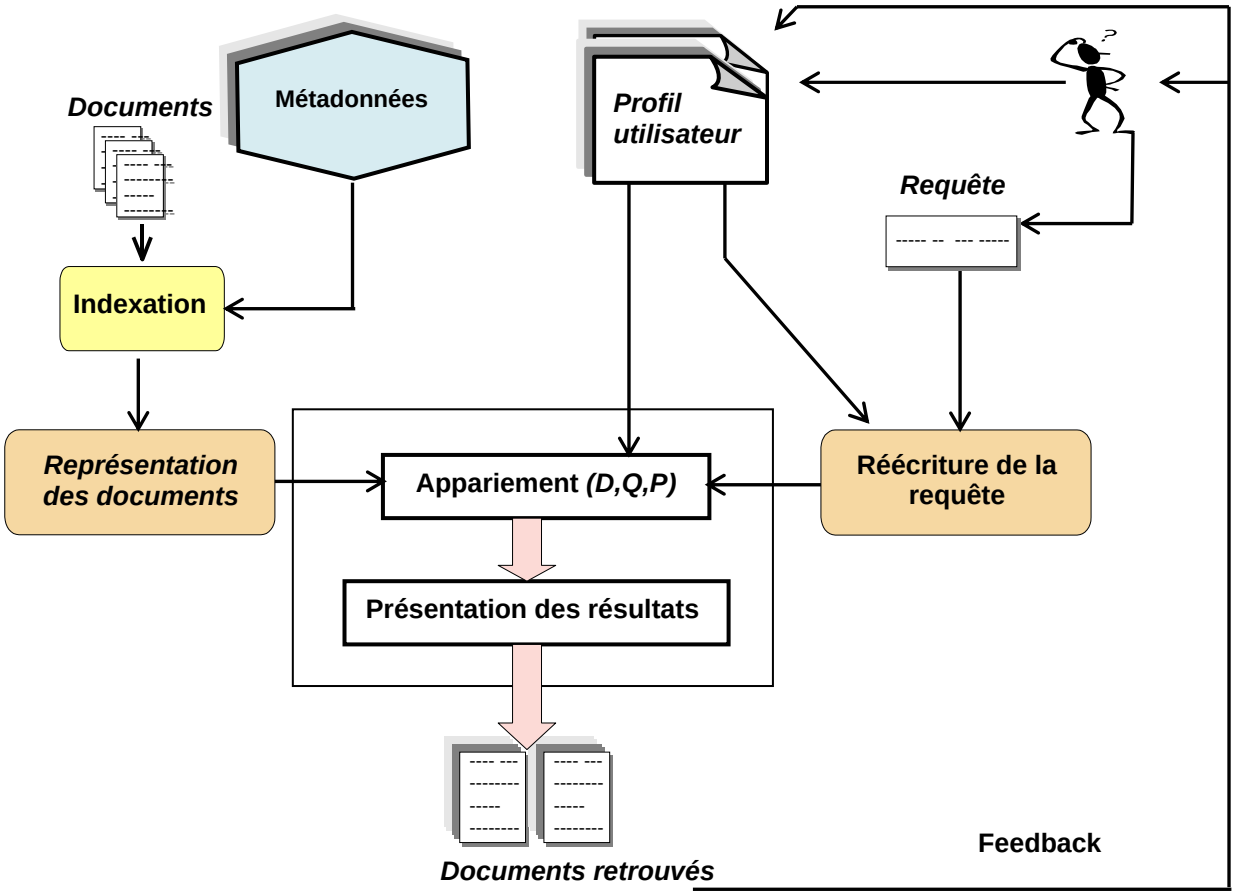

Figure 2.1 : *Architecture générale d'un SRIP.*

II.2 Notion de profil

Dans le but de personnaliser l'information, le SRIP doit disposer des éléments clés ayant une incidence concrète sur la recherche en cours. Les éléments en question représentent les données contenues dans le profil utilisateur et les métadonnées des documents. En effet, l'utilisateur et le document ne sont pas assimilés seulement à des descripteurs exprimés à l'aide de mots comme c'est le cas dans les SRI classiques. Ils possèdent tous deux des caractéristiques propres.

II.2.1 Le profil utilisateur

L'utilisateur est un élément clé dans la personnalisation en RI. Il représente le noyau central d'un SRIP : il est la source, le déclencheur d'une RI et le seul à valider le résultat de cette recherche.

On appelle **profil utilisateur** *toute structure qui permet de* ***modélise****r 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.

Il convient de distinguer la notion de *profil* de la notion de *requête*. Un profil est 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.

Pour modéliser le profil, les questions à poser sont : comment les préférences d'utilisateur peuvent être obtenues et comment seront elles structurées [Goecks 00] ?
En effet, le processus de modélisation du profil utilisateur requiert deux étapes [Amato 99]. Il faut déterminer en premier le « Quoi » puis le « Comment » :

- Le «Quoi » : c'est la détermination de ce que doit représenter le contenu du profil.
  Quelles sont les informations pertinentes représentant au mieux les centres d'intérêts et les besoins de l'utilisateur ?
- Le « Comment » : c'est la détermination de la structure du profil et la/les technique(s) à utiliser pour la construction de ce profil.

II.2.2 Le profil de document

Le document, élément central d'un SRI, est un objet complexe sans cesse en évolution car il est lié aux développements des technologies de la communication.
La finalité d'un document est de transmettre de l'information. Mais cette information est un concept abstrait, porteur de sens. Ainsi l'information d'un même document, varie d'un individu à un autre selon la perception individuelle de chacun.

En outre, le sens d'un document est donné, non seulement par son contenu mais aussi par sa structure
L'idée majoritaire aujourd'hui, est de renforcer la représentation de l'information du document en utilisant l'information sur l'information du document: les *méta informations*. On parle aujourd'hui de **métadonnées** et d'indexation « non thématique » en opposition à l'indexation « thématique » ou « conceptuelle » décrivant la sémantique de l'information (les descripteurs par exemple) [Ben Abdallah 97].

L'objectif principal des métadonnées est de décrire, d'identifier et de définir une ressource. Toutefois, c'est dans leur rôle que repose le véritable intérêt des métadonnées: faciliter la RI [Perenon 00].

Les parties de document ont un usage différencié à priori suivant le besoin de l'utilisateur. Dans la majorité des cas, un document a une *structure générale* qui forme une unité, car il est construit pour faire passer un message. Cette unité matérielle et intellectuelle est le résultat d'un lien parfaitement établi entre ses différentes parties, celles-ci pouvant former à leur tour des unités indépendantes remplissant une fonction bien déterminée.

Un SRIP gère ainsi, en plus des documents entiers, des parties de documents appelées « unités documentaires » afin de personnaliser l'information à présenter aux utilisateurs .L'éclatement du document en unités documentaires permet, tout en préservant l'unité globale du document (le lien entre l'unité documentaire), de présenter à l'usager une information plus affinée et plus facile à saisir.

II.3 Accès personnalisé à l'information

Un SRIP ne se limite pas seulement à modéliser les caractéristiques des utilisateurs en des profils ou à représenter les collections de documents par des métadonnées. En outre, il doit être capable de déduire à partir de ces profils, l'intention de l'utilisateur lorsqu'il effectue sa recherche. En effet, comme la recherche d'information s'inscrit dans le cadre d'une activité ponctuelle de l'utilisateur, pouvoir déterminer ce but permet au SRIP de mieux satisfaire l'utilisateur.
Outre la connaissance des centres d'intérêts de l'utilisateur, la connaissance de ses préférences particulièrement celles liées à la présentation des résultats, est une phase importante que doit inclure un SRIP, pour offrir un accès personnalisé à l'utilisateur, et augmenter ainsi la portée du service de personnalisation offert par le SRIP.

On distingue donc deux objectifs d'un SRIP : connaître le *but de la recherche* de l'utilisateur pour mieux le satisfaire et adapter la *présentation de l'information* pertinente en fonction des exigences de chaque utilisateur.

II.3.1 Le but de la recherche

L'utilisateur initie la recherche dans un but bien précis, poussé par un besoin en informations. Ce but est fortement lié à une activité ponctuelle de l'utilisateur.
S'adapter à ce but (et par conséquent, l'adaptation à la tâche de l'utilisateur) permet de fournir l'information qui satisfait le mieux l'utilisateur.
Lorsque le système connaît le contexte dans lequel s'inscrit la recherche et quel est le besoin informationnel de l'utilisateur, il est plus apte à fournir l'information pertinente liée à la recherche en cours.

Il existe deux types de besoins utilisateur : le besoin à court terme et le besoin à long terme. Ils dépendent tous deux intégralement de l'utilisateur :

- Le *besoin à court terme* représente le besoin ad hoc et occasionnel de l'utilisateur. Il dépend de l'activité (tâche) actuelle de l'utilisateur c'est-à-dire la tâche pour laquelle l'utilisateur initie la recherche. C'est l'équivalent de ce que l'on appelle la requête en RI classique.

- Le *besoin à long terme* a une persistance plus longue et représente les centres d'intérêts de l'utilisateur et est généralement stable dans le temps.

En fait, les besoins en informations peuvent varier, en ce qui concerne leurs types, leurs contenus et leurs durées d'un utilisateur à un autre. En plus d'être hétérogènes pour des utilisateurs différents, les besoins en informations sont également hétérogènes pour un même utilisateur. Ceci est dû à la nature même de l'utilisateur qui peut avoir des préférences et des centres d'intérêts variés [Amato 99].

II.3.2 Représentation des résultats de recherche

Le but d'un SRIP n'est pas de retourner que des documents dont le contenu est pertinent, mais également d'adapter et de personnaliser l'interface de restitution et d'affichage de ces documents selon le profil de chaque utilisateur.
La personnalisation des SRI doit donc être mise en application à deux niveaux : au niveau de l'interface par laquelle l'utilisateur interagit avec le système, et au niveau du contenu informationnel [Pednault 00].

Personnaliser l'interface revient à adapter l'affichage des résultats aux préférences et aux capacités des utilisateurs. Cette adaptation s'articule autour deux aspects : l'aspect technologique et l'aspect humain.
Du côté technologique, la représentation de l'information implique l'emploi de structures de données. Ces structures de données appropriées sont exigées non seulement, pour soutenir les divers médias utilisés mais également, pour coder des données spécifiques de l'utilisateur requises pour déterminer le contexte et les buts courants de l'utilisateur.

Du côté humain, cela exige que les systèmes de personnalisation s'adaptent à la manière de communiquer, de recevoir et d'organiser l'information selon les souhaits de l'utilisateur. Le système doit être capable de définir ces besoins de représentation afin de réaliser le niveau d'adaptation et de représentation désirés.

III Mise en œuvre d'un SRIP

Les deux grandes étapes du développement d'un SRIP sont la modélisation de l'utilisateur et celle du document.

III.1 Modélisation de l'utilisateur

Modéliser l'utilisateur, ce qu'il est, ses centres d'intérêts et ses préférences est une tâche primordiale pour les concepteurs des SRIP.

Pour modéliser l'utilisateur il faut définir en premier la *structure* de son profil qui permet non seulement, de stocker les informations le concernant mais aussi de les exploiter d'une manière optimale. En second, il faut déterminer les techniques de construction et de mise à jour de ce profil. Le type d'approche adoptée par le SRIP détermine fortement l'efficacité du système.

Plusieurs approches et techniques on été développées pour modéliser l'utilisateur ; cependant elles diffèrent dans la manière de représenter, de construire et de mettre à jour le profil. Le contenu informationnel du profil dépend aussi fortement de l'application. Le plus souvent c'est les données exploitées par le système qui déterminent le contenu du profil.

III.1.1 Représentation du profil utilisateur

Comme le contenu du profil dépend fortement de l'application qui l'exploite, on trouve dans la littérature trois principales approches pour représenter et structurer les profils des utilisateurs : représentation vectorielle, hiérarchique et multidimensionnelle.

III.1.1.1 Représentation vectorielle

Ce type de représentation s'appuie généralement sur le modèle vectoriel [Salton 71]. Le contenu du profil est constitué d'un ou de plusieurs vecteurs définis dans un espace de termes. Ces termes sont obtenus à partir de plusieurs sources d'informations concernant l'utilisateur.
Les coordonnées des vecteurs correspondent aux poids associés aux termes retenus dans le profil. On peut citer comme exemple des systèmes : Alipes, WebMate et Surfagent [Somlo 03].
L'utilisation de plusieurs vecteurs correspond à deux préoccupations : pouvoir prendre en compte des centres d'intérêt multiples et gérer leur évolution dans le temps.
Cette représentation apporte l'avantage de la simplicité de mise en œuvre. Néanmoins, même si ces systèmes prennent en considération des centres d'intérêts multiples en utilisant plusieurs vecteurs, cette représentation manque de structuration. Cette représentation ne facilite ni

l'interprétation ni la prise en compte des différents niveaux de généralités caractérisant l'utilisateur [Bottraud 04].

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. Il faut donc modéliser le profil utilisateur de façon à prendre en considération l'ensemble des paramètres représentant l'utilisateur.

III.1.1.2 Représentation hiérarchique

De nombreuses approches ont été suivies pour améliorer d'avantage les performances des SRIP en structurant mieux les profils utilisateurs. La construction d'une hiérarchie de concepts ou d'une ontologie personnelle, plutôt qu'un ensemble de domaines indépendants (suivant une direction proposée dans un contexte plus général par Huhn [Huhn 99]) offre une alternative intéressante à l'approche précédente.
Dans cette approche, la modélisation de l'utilisateur est fondée sur l'élaboration d'une ontologie personnelle. L'ensemble des caractéristiques de l'utilisateur est organisé dans une structure hiérarchique de concepts (catégories) où chaque catégorie représente la connaissance d'un domaine d'intérêt de l'utilisateur.
Le SRIP s'appuie sur la sélection dans une ontologie générale de nœuds estimés correspondre aux intérêts de l'utilisateur.
Ainsi, le rapport de généralisation /spécification existant naturellement dans ce genre de structure permet d'avoir une représentation plus réaliste du profil utilisateur.

Le premier à avoir utilisé une telle structure fut Pretschner [Pretschner 99] dans le système OBIWAN. Il a proposé un modèle innovant pour la construction du profil utilisateur. Il s'appuie sur l'ontologie publique de Magellan[1] qui est composée d'approximativement 4.400 nœuds de concepts. Semblable à ce travail, on peut citer le système SmartPush [Kurki 99] [Gauch 03].

Bien que la représentation de ce profil d'utilisateur soit innovatrice, ces travaux ne se servent pas des caractéristiques de la structure hiérarchique (par exemple pour dédoubler ou fusionner des nœuds dans le profil d'utilisateur) pour capturer la dynamique des changements. De plus, la sémantique générale de cette hiérarchie n'est pas formellement indiquée; dans la plupart des cas, ils correspondent à une relation de généralisation/spécialisation.

[1] http://www. magellan.excite.com

III.1.1.3 Représentation multidimensionnelle

La représentation multidimensionnelle du profil s'inscrit dans une réflexion globale sur la personnalisation de l'information. En effet, le profil utilisateur est un élément clé dans le processus de recherche, la modélisation de l'utilisateur doit pouvoir capturer toutes les dimensions qui représentent l'utilisateur.

Différents travaux ont abordé cet aspect sans le couvrir dans son ensemble. Ainsi, les propositions de standards P3P [W3C 04] 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 [Amato 99] consiste à représenter le contenu du profil utilisateur par un modèle structuré de dimensions (ou catégories) prédéfinis. C'est la première approche où les informations sont structurées et qui offre un modèle général.
Le modèle de profil contient cinq catégories :

1- catégorie de données personnelles,
2- catégorie de données de la source,
3- catégorie de données de livraison,
4- catégorie de données de comportement,
5- catégorie de données de sécurité.

La première catégorie *Données personnelles* contient toutes les informations concernant l'identité de l'utilisateur.

La deuxième « *Données collectées* » contient les informations nécessaires pour décrire les préférences et restrictions sur les documents. Elle est divisée en trois sous catégories : *contenu* (des informations sur le sujet du document, la langue, etc.) *structure,* (format, type, date de publication, dimensions, etc.), *source* (provenance, auteurs, éditeurs, etc.).

Dans la catégorie « *Données de livraison* », on trouve les informations sur la manière de transmettre des résultats à l'utilisateur. Ces informations sont regroupées selon deux sous catégories : *moyen* (mode de livraison par exemple email, fax téléphone, etc.) et *moment* (contient des informations temporelles sur le moment de livraison comme lors d'un changement, vers midi, entre 9h et 9h15, etc.).

Dans la catégorie « *Données de comportement* » se trouvent des enregistrements sur les interactions de l'utilisateur avec le système (URLs des pages visitées, documents lus et pertinence, etc.).

Enfin dans, la catégorie *Données de sécurité,* des informations sont données sur les conditions d'accès aux données du profil.

L'auteur a proposé ce modèle dans le cadre du développement d'un service avancé de librairie digitale (recherche et de livraison personnalisé de l'information sur le Web) : le système EURO*gather service*.

Poursuivant la classification de [Amato 99], Kostadinov [Kostadinov 03] propose un ensemble de dimensions ouvertes, capables d'accueillir la plupart des informations caractérisant un profil.
Il distingue principalement huit dimensions :

1. les données personnelles,
2. le centre d'intérêt,
3. l'ontologie du domaine,
4. la qualité attendue des résultats délivrés,
5. la customisation,
6. la sécurité et la confidentialité,
7. le retour de préférences (feedback),
8. les informations diverses.

Ces classes de données sont brièvement décrites 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 (age, 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'intérêts 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 de 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.
Les dimensions et sous dimensions définissant un profil ne sont pas indépendantes les unes des autres; elles peuvent être liées par des associations sémantiques qui caractérisent leurs dépendances ou leurs corrélations.

III.1.2 Construction du profil

Un des problèmes les plus difficiles à résoudre dans la personnalisation consiste à trouver des techniques optimales de construction et de mise à jour des paramètres du profil d'un utilisateur.
En effet, la complexité des techniques adoptées par les systèmes pour l'acquisition des éléments du profil de l'utilisateur, dépend fortement du modèle de représentation associé. Plus le modèle de représentation est simple, plus la technique de construction et de mise à jour est simple, et inversement.

Les SRIP construisent le profil utilisateur à partir de sources d'informations. Ces sources d'informations sont acquises par différentes approches :

- des approches manuelles, en se basant sur les informations fournies directement par l'utilisateur,
- des approches automatiques (ou semi-automatiques) par apprentissage à partir des documents consultés et/ou du comportement de l'utilisateur lors de différentes sessions de recherche
- des approches fondées sur l'identification de groupes d'utilisateurs et la détermination des caractéristiques clés de chaque groupe.

Les SRIP appliquent généralement les mêmes techniques pour collecter ces sources d'informations que les SRI adaptatifs : système de filtrage collaboratif, explicite ou implicite ou les techniques de jugement de pertinence de l'utilisateur (feedback). Néanmoins, de nouvelles techniques sont développées afin d'améliorer les techniques utilisées habituellement par les SRI adaptatifs.

Les SRI ont recours à différents indicateurs pour extraire et structurer les connaissances représentatives de l'utilisateur et analyser son comportement de navigation à partir d'ensembles [Goeckst 00] [Claypool 01] tels que : les mouvements et clicks de souris, les temps passés sur la page, le nombre de liens activés, etc. Aucun feedback d'utilisateur n'est nécessaire.
Après avoir regroupé ces sources de données, le système doit être capable de les analyser et d'en déduire le profil de l'utilisateur et, également de les stocker dans la structure spécifiée qu'il utilise. Il est à noter que les techniques de construction du profil dépendent fortement du modèle de représentation. On en distingue trois grandes approches :

III.1.2.1 Analyse statistique des termes

L'analyse statistique des mots clés est la méthode la plus utilisée parce qu'elle est basée sur des techniques bien comprises d'extraction de mots clés.
L'idée principale consiste à analyser le contenu des documents jugés pertinents par l'utilisateur et d'en extraire des mots clés significatifs qui décrivent son contenu. Dans certains cas, le rajout d'un poids exprime l'importance de chaque terme et est souvent associé à la fréquence d'apparition du terme.

Les inconvénients de cette approche se trouvent dans le fait qu'elle ne peut être appliquée que sur des éléments textuels et que les mots sont analysés en isolation avec le reste du document ce qui entraîne une perte d'information contextuelle pouvant dégrader l'exactitude des données du profil. Une fois les mots clés extraits.

III.1.2.2 Techniques d'apprentissage

Le principe de base de ces techniques est l'étude du comportement de l'utilisateur et la classification de ses caractéristiques. Elles utilisent généralement des algorithmes de classification. Ces algorithmes extraient les termes à partir de ces différentes sources et les regroupent en des classes. Chaque classe indique un domaine particulier des centres d'intérêts de l'utilisateur. L'ajustement du poids des termes fait de plus en plus appel à des techniques d'apprentissage, comme des réseaux de neurones [Shavlik 99], des probabilités Bayésiennes [Pohl 99], des algorithmes à base de règles [Bloedorn 96] [Krulwich 97].

L'avantage de l'approche est la fraîcheur et l'exactitude des données dérivées. Grâce au suivi de l'utilisateur cette approche permet la mise à jour de son profil. Elles sont largement utilisées dans les systèmes qui représentent le profil utilisateur par des vecteurs de termes

La construction d'une ontologie personnelle, pour représenter le profil utilisateur, peut également être réalisée à partir des techniques d'élaboration automatique des thesaurus à partir de l'analyse de collections de documents utilisant les algorithmes d'apprentissage.
Chaque nœud de la structure hiérarchique est associé à un ensemble de documents qui sont employés pour représenter la teneur de ce nœud. Tous les documents associés à un nœud (généralement 10 documents par nœud) sont fusionnés dans un super document.

L'inconvénient de cette approche se trouve dans la complexité des algorithmes utilisés qui nécessitent beaucoup de temps.

III.1.2.3 Concept de la vie artificielle

L'utilisation la théorie de la vie artificielle pour construire et mettre à jour le contenu du profil, est une approche assez novatrice [Chen 01].
Le profil d'utilisateur, appelé dans ce cas « la vue personnelle », est décrit par un ensemble d'éléments représentant ces centres d'intérêts. Chaque élément dans la vue personnelle est décrit par un vecteur Vi de mot-clé et une valeur Ei d'énergie.
Tandis que Vi, qui contient un ensemble de mots-clés et de leurs poids correspondants, décrit la teneur des intérêts d'utilisateur, Ei indique l'importance de la catégorie.
L'énergie Ei 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 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.
Basé sur les valeurs d'énergie des catégories, la structure de la vue personnelle peut être modulée pendant que les intérêts des utilisateurs changent.
Ce type de technique s'appuie, également sur les méthodes de classification pour établir la vue personnelle. Les intérêts de l'utilisateur sont dépistés et classés par catégories; le système se base sur une hiérarchie prédéfinie de catégories, appelée « la vue du monde », comme une hiérarchie de catégories supérieure à la vue personnelle.

III.2 Modélisation des documents

Les documents sont porteurs d'informations. Ils servent à transmettre une connaissance (un savoir, un message informationnel) d'un producteur vers un consommateur (l'utilisateur).
Cette capacité de transmission est obtenue grâce à la mise sous une forme persistante de cette information. La perception individuelle qu'a chaque utilisateur de la sémantique d'un même document, rend difficile la construction d'une base commune de classement. En effet, il n'existe pas de représentation unique et donc commune de l'information. Cependant, l'une des étapes primordiales dans un processus de RI c'est l'indexation des corpus documentaires.

Néanmoins, cette indexation ne permet pas de déterminer très précisément les informations contenues par le document. Il est donc difficile pour un SRI de vérifier la pertinence réelle de chaque document. Pour pallier aux limites de l'indexation et pour avoir une meilleure connaissance du corpus documentaire, les SRIP tentent de décrire les documents par des critères externes à leurs contenus. Ainsi, l'idée est de renforcer la représentation de l'information du document en utilisant l'information sur l'information du document : les méta informations (métadonnées).

Le terme de métadonnées est utilisé pour définir l'ensemble des informations techniques et descriptives ajoutées aux documents pour mieux les qualifier. Pour que ces données soient utilisables par d'autres, elles doivent s'inscrire dans des modèles largement reconnus par les acteurs du Web.
Plusieurs organismes de standardisation ont donc proposé et publié des schémas de métadonnées susceptibles d'être utilisés par le plus grand nombre.
Le schéma de métadonnées le plus utilisé est proposé par l'organisation Dublin Core Metadata Initiative (DCMI) ; on l'appelle le plus souvent le Dublin Core. Il standardise l'utilisation d'une quinzaine de champs descriptifs. URC (Uniform Ressources Characteristic or Uniform Resources Citation), USMARC format, TEI (Text Encoding and Interchange) sont quelques exemples.

- **Les éléments Dublin Core**

Dublin Core précise la sémantique de 15 métadonnées [Dublin 04]:

1- Title	Titre (*le titre de la ressource*)
2- Creator	Auteur ou créateur (l'auteur *de la ressource*)
3- Subject	Sujet et mots-clés (*description du sujet de la ressource*)
4- Description	Description (*description, résumé de la ressource*)
5- Publisher	Éditeur (*personne ou institution en charge de la diffusion de la ressource*)
6- Contributor	Autres auteurs (*co-auteurs*)
7- Date	Date (*Date de publication*)
8- Type	Type de ressource (*catégorie de la ressource : roman, poème, rapport technique, dissertation, dictionnaire*)
9- Format	Format (*format de données de la ressource, par exemple HTML*)
10- Identifier	Identifiant unique de ressource (*URL, ISBN*, ...)
11- Source	Source (*source dont le document a été dérivé, ex titre de la revue*)
12- Language	Langue (*langue de la ressource, par exemple fr*)
13- Relation	Relation (*rapports avec d'autres ressources, IsPartOf, IsVersionOf,...)*
14- Coverage	Couverture (couverture *géographique ou temporelle de la ressource*)
15- Rights	Gestion de droits (*information au sujet de la propriété du document*)

Outre les propriétés classiques de description des documents (titre, auteur, pays, année, environnement de production, environnement de diffusion) défini par le Dublin Core, le système Profil-Doc [Perenon 00] caractérise les unités documentaires selon trois propriétés :

Type	Résumé Table des matières Introduction Description de contexte Description de thème Environnement Expérimentation Résultats Discussion Méthode Conclusion Bibliographie
Forme discursive	Descriptif, narratif, argumentatif, discours rapporté
Style de présentation	Littéraire Littéraire avec données numériques Données numériques Calcul Représentation

Cependant, comme le système Profil-Doc est un système dédié à la recherche de documents scientifiques et techniques, la définition de ces modalités a été proposée à partir d'un questionnaire effectué auprès de chercheurs en SIC, sciences pharmaceutiques et sciences physiques.

Toutes ces stratégies ont deux caractéristiques : elles portent sur des critères (la forme, le support, le style, ...) autres que le contenu du document, elles sont très fortement individualisées et permettent une personnalisation de la recherche
En effet, ces propriétés nous permettront de sélectionner un corpus "personnalisé" suivant les caractéristiques de l'utilisateur, corpus sur lequel portera la question.

Les documents sont donc découpés suivant la structure logique du document. Le SRIP utilise ces différentes structures pour décrire les documents en unités documentaires. Chacune des unités est alors accessible par des index bien sûr, mais aussi par ses propriétés. Le découpage est basé sur la fonction remplie par ces parties du document et non sur leur contenu

IV Conclusion

Les problématiques liées à l’hétérogénéité des applications et à la diversité des exigences des utilisateurs, font émerger la personnalisation comme une approche capitale dans la conception et le développement des SRI du futur.
La personnalisation de l’information a pour but de fournir une information pertinente, qui correspond exactement aux besoins de l’utilisateur.
Répondre aux besoins en information des utilisateurs d’une manière personnelle, ne peut se faire sans inclure l’utilisateur dans le processus de RI. Inclure l’utilisateur dans le processus de RI implique la représentation de ce dernier dans un modèle ou par une structure qui permet son exploitation par le SRI.

Plusieurs axes de recherche et de nombreux industriels se sont appliqués à développer des systèmes de recherche et d’accès à l'information, adaptés aux besoins spécifiques de l’utilisateur et qui fournissent l’information désirée en fonction de l’utilisateur qui la recherche.
Chacun des ces systèmes de personnalisation propose sa propre approche de construction des profils utilisateurs. Comme le contenu et la structure du profil dépendent fortement de l’application, le problème de capture des paramètres du profil reste encore posé. Le fait que les utilisateurs ont des centres d’intérêts et une demande d’informations différentes et qu’en plus, ils ont souvent des idées floues sur leurs préférences, complique l’extraction des caractéristiques pertinentes et nécessite un processus de découverte de leurs préférences.

Un système de personnalisation devrait automatiquement découvrir les différents centres d’intérêts de l’utilisateur avec une implication minimale de ce dernier, afin d'identifier ce dernier de manière spécifique. Le système doit être capable également de détecter les changements des centres d’intérêts pour effectuer la mise à jour du profil, pour qu'il évolue en fonction de l'évolution de l'utilisateur.

L’élaboration d’un modèle de profil générique semble être une meilleure approche pour pouvoir caractériser entièrement l’utilisateur, ce qu’il est, ses préférences, son environnement et tout ce qui le représente en tant qu’individu à part entière. L’utilisateur ne devient plus qu’un simple acteur dans le processus de recherche mais, une entité essentielle et indivisible du processus.

Chapitre 3

Définition et intégration du profil utilisateur dans le processus de RI pour un accès personnalisé à l'information

I Problématique et objectifs

Le but d'un SRI est de fournir l'information pertinente aux utilisateurs. Les efforts fournis par la communauté de RI durant ces derniers années pour améliorer les performances des SRI ont permis de faire évoluer les SRI classiques vers les SRI adaptatifs, puis vers les SRI personnalisés.

Idéalement, un SRIP doit être capable de découvrir, d'extraire, de structurer, de stocker puis d'exploiter et de mettre à jour toutes les caractéristiques pertinentes représentatives de l'utilisateur, avec une implication minimale de ce dernier. Le SRIP gère les corpus documentaires de façon à pouvoir cibler l'information pertinente en fonction des besoins informationnels et du profil de chaque utilisateur. Il doit donc pouvoir manipuler des descripteurs de documents reflétant au mieux leur contenu sémantique.

Cependant, malgré le succès des approches utilisées pour personnaliser les SRI, ils présentent certains inconvénients concernant l'initialisation, la représentation, l'exploitation et l'évolution du profil utilisateur :

- La construction du modèle du profil est fortement guidée par l'utilisateur et le temps qu'il y consacre.
- Les caractéristiques utilisées pour représenter les préférences et les centres d'intérêts de l'utilisateur sont peu nombreuses et ne couvrent pas l'ensemble de ses connaissances.
- Le contexte dans lequel la recherche est effectuée est rarement utilisé pour comprendre la requête et la situer par rapport au profil utilisateur.
- L'évolution du profil reste très difficile à maîtriser à cause des différents changements associés aux utilisateurs.

Il est donc clair que si l'on veut offrir un accès personnalisé à l'information en fonction de chaque utilisateur, il faut pouvoir modéliser l'ensemble des paramètres caractérisant l'utilisateur et le document. Il faut alors définir des modèles pour représenter le profil de l'utilisateur et le profil du document ainsi que des stratégies pour les inclure dans le processus de RI.

L'objectif de notre travail s'inscrit dans le cadre d'un projet de développement d'un SRIP permettant un accès personnalisé à l'information pertinente en fonction de chaque utilisateur. Le SRIP intégrera fortement la composante utilisateur tout au long de la chaîne d'accès à l'information pertinente.

Pour développer un SRIP performant il faut avoir une vision globale de ce qu'est « l'accès personnalisé à l'information » afin de cerner tous les paramètres déterminants de la personnalisation.

Les problèmes liés à la mise en œuvre d'un SRIP incluant le profil utilisateur sont :

- Quel est le modèle à utiliser pour représenter le profil utilisateur ?
- Que doivent contenir le profil ?
- Comment inclure le profil dans le processus de recherche ?
- Comment faire évoluer les profils ?

Face à ces problèmes, nous nous sommes particulièrement intéressées dans le cadre de ce présent travaille à effectuer une exploration aussi large que possible de la notion de profil. Une réflexion sur l'utilisabilité et l'exploitation du profil utilisateur dans le processus de recherche nous a conduit vers une catégorisation du contenu du profil utilisateur. La proposition d'une architecture (modulable) permettant de supporter l'ensemble des processus mis en œuvre pour effectuer la personnalisation a été également abordée dans le contexte de notre étude.

Les objectifs poursuivis durant notre étude sont les suivants :

1- Modélisation du profil utilisateur.

2- Définition l'architecture de SRIP.

3- Définition des interactions profil utilisateur – phases du processus de recherche.

Ce chapitre est décomposé en trois grandes parties. La première partie concerne la définition du profil utlisateur. On abordera dans la seconde partie les étapes suivies pour le développement du SRIP, ainsi que la définition de l'architecture du système.
Dans la dernière partie de ce chapitre, on détaillera les approches adoptées pour intégrer les profils dans le processus de recherche.

II Définition du profil utilisateur

L'étude des différentes approches utilisées pour représenter les connaissances caractérisant l'utilisateur nous a fait aboutir à la conclusion qu'il est nécessaire d'avoir une représentation multicatégories (multidimensionnelles) du profil utilisateur pour pouvoir capturer l'ensemble des paramètres le caractérisant [Amato 99] [Kostadinov 03].

En effet, l'utilisateur est caractérisé par différents niveaux de connaissances : des besoins en informations à long et à court terme, des préférences et des centres d'intérêts variés, des exigences de représentation, une activité professionnelle etc. L'ensemble de ces caractéristiques représente son profil. Rappelons à cet effet qu'un profil utilisateur correspond à toute représentation permettant de caractériser l'utilisateur dans un processus de recherche. La représentation est un modèle formel supporté par des structures de données, qui permet l'exploitation optimale des données du profil.

La définition d'un modèle formel généralisable à toute application semble être une tâche très difficile du fait de la diversité des caractéristiques des utilisateurs et des facteurs peu paramétrables le représentant. Cependant, il est possible de catégoriser le contenu des profils en fonction du rôle qu'ils vont avoir dans le processus de recherche personnalisé.

L'approche abordée pour modéliser l'utilisateur est basée sur l'impact que peut avoir le contenu du profil utilisateur sur le processus de recherche pour établir une catégorisation de ce contenu. En effet, les phases du processus de recherche sont différentes et les approches pour personnaliser ces phases sont également différentes. Il est donc évident que l'on ne va pas utiliser le profil utilisateur de la même manière dans tout le processus. Chaque phase du processus de RI a besoin de certains types de données du profil pour effectuer la personnalisation de l'accès à l'information. En d'autres termes, chaque phase du processus requiert une ou plusieurs parties spécifiques du profil utilisateur.

II.1 Contenu du profil utilisateur

A partir du processus du SRIP, on regroupe dans une même dimension l'ensemble des données susceptibles d'avoir un rôle similaire dans les phases du processus de recherche. On distingue donc trois grandes dimensions représentant l'utilisateur : les données de préférences concernant la recherche effective, ses données personnelles qui permettent de l'identifier, ainsi que les données du l'environnement de travail de l'utilisateur.

On propose alors la catégorisation suivante :
- catégorie des préférences,
- catégorie des données personnelles,
- catégorie des données de l'environnement.

Cette catégorisation n'a pas la prétention de cibler tous les paramètres caractérisant l'utilisateur; il est possible d'étendre le profil utilisateur en rajoutant d'autres catégories liées spécifiquement à une application. Seulement il faut souligner que l'ajout d'une catégorisation du contenu du profil doit se faire en prenant comme principal critère de choix le rôle de ces données dans le processus de personnalisation de l'application.

II.1.1 Catégorie des préférences

Dans cette catégorie on regroupe l'ensemble des informations correspondant aux préférences et centres d'intérêts de l'utilisateur. C'est la catégorie la plus importante dans le profil utilisateur. Elle a une incidence directe sur le processus de recherche, puisque c'est à partir de ces informations que le système va pouvoir retourner l'information pertinente correspondante aux besoins de l'utilisateur.

Cette catégorie peut être décomposée en deux sous catégories car on distingue en fait deux types d'informations : **les informations représentant les centres d'intérêts de l'utilisateur** et **les informations représentant les préférences de recherche de l'utilisateur**. Ces deux sous catégories sont nommées respectivement : « Domaine d'intérêts » et « Préférence de recherche ».

a) ***Domaine d'intérêts***

Cette sous catégorie regroupe les informations caractérisant les centres d'intérêts de l'utilisateur. Ces informations sont relativement stables dans le temps et représentent les besoins à long terme de l'utilisateur. Elle contient la description des documents consultés par l'utilisateur lors d'une ou des sessions de recherche.
Ces documents représentant un intérêt particulier pour l'utilisateur et n'ont cependant pas le même le degré d'importance. L'ensemble des documents sera donc regroupé dans une hiérarchie de classes représentant chacune un domaine d'intérêt de l'utilisateur. Ces domaines peuvent directement être spécifiés par l'utilisateur.

b) Préférences de recherche

Cette sous catégorie regroupe l'ensemble des préférences de l'utilisateur lié aux informations recherchées. On y distingue trois types de préférences : celles liées au processus de recherche, celles liées aux données recherchées ainsi que les préférences de restitution des documents. Chacun de ces types a un nombre déterminé d'attributs qui seront spécifiés par l'utilisateur.

- Les préférences liées au processus de recherche sont caractérisées par les attributs : temps de réponse et type de recherche. L'attribut « temps de réponse » correspond au temps maximal souhaité par l'utilisateur. L'attribut « type de recherche » correspond au type de recherche que l'utilisateur souhaite effectuer.

- Les préférences concernant les données recherchées se décomposent en deux types de préférences : les préférences concernant les données elles-mêmes et les préférences concernant la structure des documents. Les attributs liés à ces préférences sont : type_source, type_fournisseur ; langue, format et date_création.

- Les préférences de restitution sont communément appelées « préférences de customisation ». Elles regroupent un ensemble de paramètres liés aux spécificités requises par l'utilisateur pour livrer et afficher l'information pertinente retrouvée. Ces préférences de customisation se décomposent en deux paramètres : les préférences de livraison et les préférences de mise en page. Chacun de ces paramètres est caractérisé par des attributs spécifiques.

Pour le paramètre de « Livraison », le type de format souhaité pour les documents restitués (PDF, PS, Doc, …), le moyen de restitution de ces documents (page web, email, fax, mobile,…) ainsi que le moment de restitution (date exacte ou intervalle de temps).

Pour le paramètre « Mise en page », les spécificités d'affichage (nombre de document par page web, partie ou tout le document, …) et les spécificités de présentation (police des caractères ; couleurs des liens, couleur,…).

L'ensemble du contenu de la catégorie des préférences peut être schématisé comme suit :

Catégorie des préférences→ (Domaines d'intérêt, Préférences de recherche)
Domaine d'intérêt →(Topic)
Préférences de recherche →(Processus, Données, Customisation)
Processus →{temps de réponse}
Données →(Contenu, Contenant)
Customisation →(Livraison, Mise en page)
Contenu →(Type_source, Type_fournisseur, Langue)
Contenant →(Format, Date_Création)
Livraison →(Type_format, Moyen, Moment)
Mise en page →(Spécificité_Affichage, Spécificité_Présentation)

II.1.2 Catégorie des données personnelles

Cette catégorie regroupe toutes les informations personnelles de l'utilisateur. Elle permet d'identifier chaque utilisateur de manière unique. Elle se décompose en deux sous catégories :

a) La sous catégorie « **Identité** » regroupe les données d'identification de l'utilisateur comme son nom, prénom, adresse, ainsi qu'un identifiant unique. Une catégorisation de la référence P3P déjà faite par le W3C peut être largement utilisée pour représenter l'utilisateur.

b) La sous catégorie « **Profession** » regroupe l'ensemble des données professionnelles de L'utilisateur. Cette sous catégorie contient comme attributs le secteur professionnel de l'utilisateur, son niveau de profession ainsi que la ou les langue(s) qu'il maîtrise.

On peut schématiser cette catégorie comme suit :

Catégorie données personnelles →(Identité, Profession)
Identité→ (Id, P3P)
Profession →(secteur, niveau profession, langue)

II.1.3 Catégorie des données de l'environnement

Cette catégorie contient les données nécessaires au système pour qu'il s'adapte à l'environnement de recherche de l'utilisateur. Ces données ont une incidence sur l'exécution du processus de recherche et la restitution en fonction des configurations logicielles et matérielles utilisées par l'utilisateur au moment de sa recherche. Elle regroupe trois sous catégories : « Logiciel », « Matériel » et « Géographique ».

a) La sous catégorie « **Logiciel** » regroupe les informations sur les logiciels installés dans l'environnement de travail de l'utilisateur. Elle concerne les données de configuration tel que le type de système d'exploitation qu'il utilise, la version du navigateur ainsi que le type d'interface s'il s'agit d'un utilisateur spécifique tels que les utilisateurs non voyants.

b) La sous catégorie « **Matériel** » regroupe les informations sur le type et les capacités du matériel utilisé lors de la recherche par l'utilisateur. Il peut s'agir d'un terminal classique tel que les PC ou d'un terminal mobile tel que les PDA, les pockets phone.

c) La sous catégorie « **Géographique** » regroupe les données concernant l'emplacement géographique d'où l'utilisateur effectue sa recherche pour spécifier les capacités de connexion de l'utilisateur.

On schématise cette catégorie comme suit :

Catégorie des données environnement →(Logiciel, Matériel, Géographique)
Logiciel → (configuration, type interface)
Matériel →(capacité, type)
Géographique →(emplacement, spécificité)

La figure 3.1 permet de visualiser l'ensemble des dimensions du profil utilisateur :

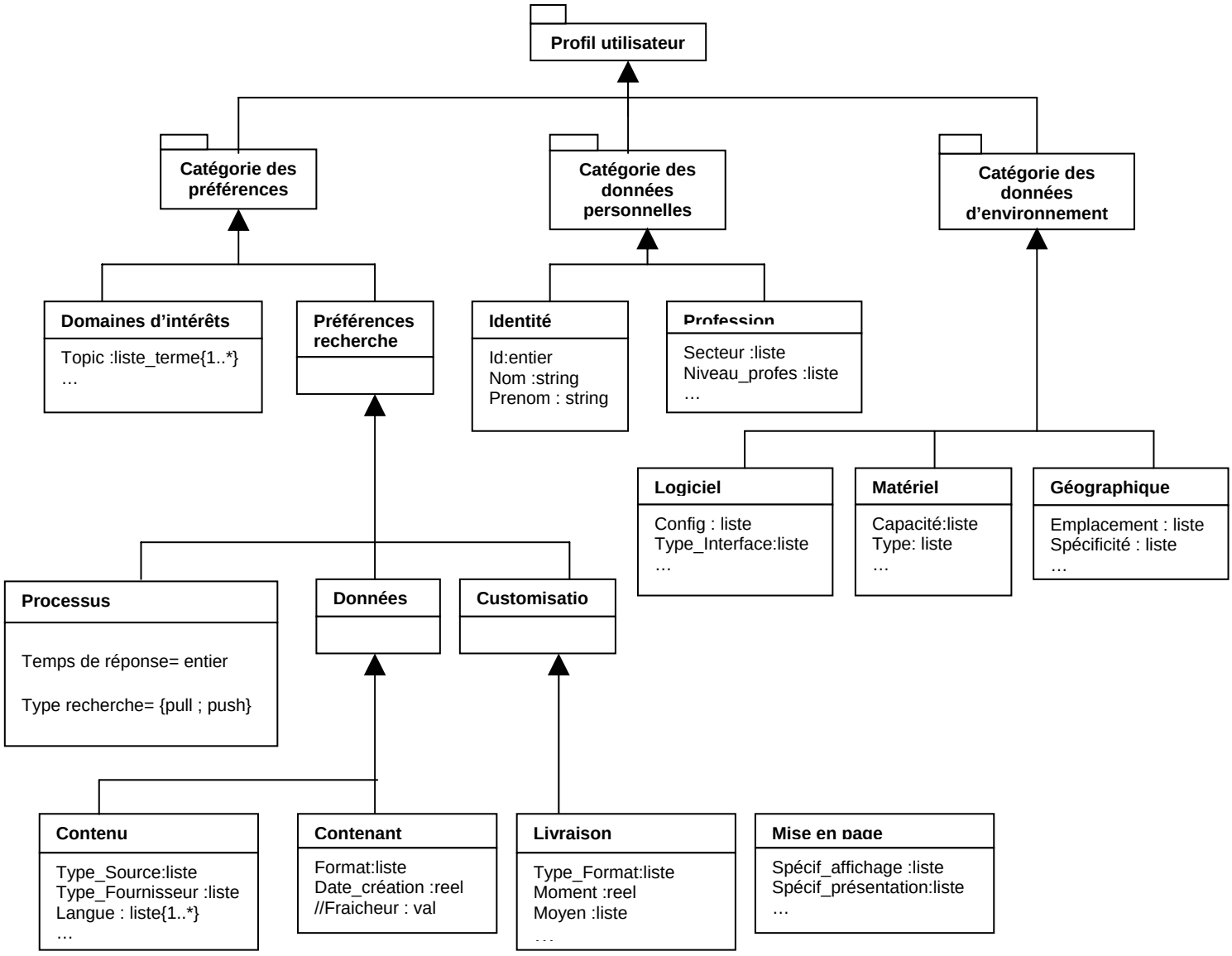

Figure 3.1 : *Représentation multidimensionnelle du Profil utilisateur.*

II.2 Représentation du profil utilisateur

Nous allons définir dans ce qui suit la représentation du profil utilisateur dans un espace vectoriel en se basant sur le modèle vectoriel de Salton [Salton 71] qui a largement prouvé son efficacité.
Le profil utilisateur est modélisé par une structure hiérarchique de catégories. Représenter une structure d'arbre dans un seul espace vectoriel va aplatir le contenu du profil au même niveau de représentation, on perdra alors la notion de dimension et de hiérarchie du profil.
Pour palier à ce problème, nous proposons de représenter le profil utilisateur dans un espace vectoriel engendré par trois espaces vectoriels représentant chacun une dimension du profil.

Le profil est donc défini par un ensemble de descripteurs vectoriels associés aux dimensions du profil : un descripteur pour chacune des trois grandes catégories du profil utilisateur. Chaque descripteur vectoriel est défini dans un espace engendré par l'ensemble des paramètres décrivant la catégorie. Nous proposons la notation suivante :

$$P = \left\{ DC_i, i = \overline{(1,3)} \right\} \quad ;$$

On défini : DC_i comme le descripteur vectoriel associé à la catégorie « *i* » , défini dans un espace vectoriel engendré par les sous catégories qui composent la catégorie « *i* ».

i = 1 : c'est la « **Catégorie des préférences** »,
i = 2 : c'est la « **Catégorie des données personnelles** »,
i = 3 : c'est la « **Catégorie des données de l'environnement** ».

Et donc $p = \left[DC_1 \quad DC_2 \quad DC_3 \right]$

Comme chaque catégorie est composée de sous catégories, les descripteurs sont eux même formés de sous descripteurs associés à chaque sous catégorie. Ces sous descripteurs sont représentés dans un espace vectoriel lui-même engendré par les attributs décrivant chaque sous catégorie.

$$DC_i = \begin{bmatrix} SC_{i1} \\ SC_{i2} \\ :: \\ SC_{iP} \\ :: \\ SC_{iM} \end{bmatrix}$$

SC_{ip} : le descripteur de la sous catégorie « *p » de la catégorie* « *i* »,

Où M : correspond au plus grand nombre maximal des sous catégories des catégories du profil.

Remarque : Comme les catégories du profil n'ont pas forcement le même nombre de sous catégories. Les sous catégories manquantes serons représentées par des vecteurs nuls.

Le profil peut donc être représenté comme suit :

$$p = \begin{bmatrix} SC_{11} & SC_{21} & SC_{31} \\ SC_{12} & SC_{22} & SC_{32} \\ \vec{0} & \vec{0} & SC_{33} \end{bmatrix}$$

On obtient une matrice (3,3), où les colonnes représentent les catégories *i* et les lignes représentent les sous catégories *p* de chaque catégorie *i*, tel que $p = \overline{(1,3)}$.

On a encore un niveau d'imbrication, puisque les sous catégories peuvent elles-mêmes être composées par d'autre catégories, nommées «Paramètre de sous catégorie». Chaque descripteur de sous catégorie est donc lui-même formé d'un ensemble de vecteurs pouvant, éventuellement, être eux même composés d'autres vecteurs.
L'ensemble de ces vecteurs est représenté dans l'espace vectoriel engendré par les termes formant les attributs des sous catégories et les paramètres de sous catégories s'ils existent.

Il est à remarquer que le degré de composition des vecteurs est relatif au niveau de profondeur de la structure du profil utilisateur.

On peut représenter cet ensemble de vecteurs imbriqués comme suit :

$$SC_{ip} = \lfloor Par_{ip1} \quad Par_{ip2} \quad ... \quad Par_{ipk} \quad ... \quad Par_{ipN} \rfloor$$

Où :

Par_{ipk} : représente le descripteur vectoriel du paramètre *k* de la sous catégorie *p* de la catégorie *i*.

N : le nombre maximal de paramètres dans toute les sous catégories du profil.

$$Par_{ipk} = \lfloor At_{ipk1} \quad At_{ipk2} \quad .... \quad At_{ipkl} \quad .... \quad At_{ipkQ} \rfloor$$

Où

At_{ipkl} : correspond au vecteur attribut *l* du descripteur de paramètre *k* de la sous catégorie *p* de la catégorie *i*.

Q : le nombre maximum d'attributs des paramètres.

Chacune des coordonnées du descripteur de paramètre correspond à un vecteur représenté dans un espace vectoriel engendré par l'ensemble des termes du domaine de définition du profil. Le contenu de chaque vecteur représente les attributs de chaque paramètre

Il est à remarquer que chaque descripteur à une interprétation spécifique selon son affiliation au type de sous catégorie et en fonction de ce qu'a été défini dans le profil utilisateur comme données.

III Mise en œuvre du SRIP

Le développement du SRIP passe par la modélisation du profil utilisateur et son intégration comme composante principale dans toutes les phases du processus de recherche.

L'intégration du profil a lieu dans toutes les phases du processus de RI : de la phase de l'expression des besoins de l'utilisateur à la présentation des résultats, en passant par la recherche effective dans les corpus documentaire et leur indexation.

Nous allons présenter dans ce qui suit une description générale du processus du SRIP. La définition d'une architecture supportant l'ensemble des phases du processus permettra de détailler dans la section suivante les interactions entre profil utilisateur et processus de recherche.

III.1 Le processus d'accès personnalisé

Les étapes du processus de personnalisation sont décrites dans ce qui suit :

III.1.1 Construction des profils

Le système construit le profil utilisateur en se basant sur le modèle multicatégories défini précédemment.
Le SRIP collecte les informations caractérisant l'utilisateur à partir de sources d'informations différentes. Puis, il les structure et les stocke dans les catégories correspondantes dans le modèle du profil utilisateur.
Ces sources d'informations sont obtenues soit explicitement, à partir de formulaire de saisies par exemple, soit implicitement à partir de techniques de scrutation des activités de l'utilisateur et de son interaction avec le système ainsi que de son environnement de travail.

III.1.2 Présélection de l'espace de recherche

Une fois les profils sont construits il faut passer à une étape primordiale pour le SRIP : la présélection de l'espace de recherche. En fait, en fonction des informations contenues dans le profil de l'utilisateur, le SRIP va regrouper l'ensemble des documents susceptibles d'être pertinents pour l'utilisateur. L'idée sous jacente est d'avoir des indexes profilés en fonction d'un utilisateur ou un groupe d'utilisateurs, dans le but de restreindre l'espace de recherche qu'aux documents qui correspondent le plus aux besoins de l'utilisateur.

III.1.3 Evaluation de la requête

Après avoir construit les profils et sélectionné un sous corpus personnalisé, le SRIP peut effectuer la rechercher effective.
L'analyse de la requête formulée va permettre de déterminer les paramètres d'instanciation du profil utilisateur. On obtient un profil de l'utilisateur circonstancié, en cela qu'il ne contient que les informations correspondantes au contexte de la recherche actuelle de l'utilisateur.

Une fois ce profil instancié, le SRIP va reformuler la requête en fonction des données contenues dans le profil utilisateur. On obtient alors une nouvelle requête plus proche du besoin réel de l'utilisateur.
Le SRIP effectuera alors un appariement entre la nouvelle requête et le corpus documentaire personnalisé pour ne sélectionner que les documents ayant une valeur de ressemblance qui dépasse un certain seuil.

III.1.4 Présentation des résultats

Avant de restituer les documents jugés pertinents pas le SRIP, une étape importante doit être réalisée. Le SRIP va adapter la restitution et l'affichage des documents en fonction des préférences de customisation contenues dans le profil de l'utilisateur. Cela permet d'avoir une valeur ajoutée au SRIP et augmenter ainsi le degré de la personnalisation.

III.1.5 Mise à jour des profils

Après présentation des résultats et à partir des réactions des utilisateurs, le SRIP va détecter les changements survenus dans le comportement de l'utilisateur et ses jugements du résultat pour effectuer la mise à jour des profils.

Pour la mise en œuvre du SRIP, il faut définir une architecture supportant les différents composants du système. Cette architecture repose sur la combinaison de différents modules ayant chacun un rôle spécifique dans le processus du SRIP.

III.2 L'architecture du SRIP

Le SRIP est construit à partir de quatre modules capables de gérer les différents composants du processus de personnalisation : le module de construction de profils, le module d'intégration de profils, le module de présentation du résultat et le module d'évolution du profil. La figure 3.2 représente le schéma général du SRIP :

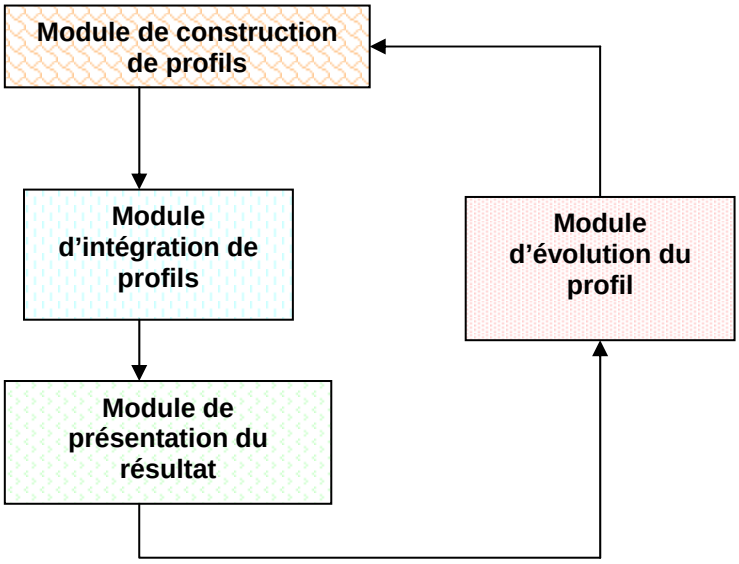

Figure 3.2 : *Architecture générale du SRIP.*

Nous présentons dans ce qui suit une description du rôle de chaque module dans le processus du SRIP.

III.2.1 Module de construction de profils

Ce module est chargé de la construction des profils à partir des modèles définis (Figure 3.3). On distingue deux sous modules : le gestionnaire du profil utilisateur et le gestionnaire du profil document.

III.2.1.1 Le gestionnaire du profil utilisateur

Il a pour rôle l'acquisition puis la structuration des connaissances représentatives de l'utilisateur à partir de différentes sources d'information. Le gestionnaire combine différentes approches pour acquérir ces données. La saisie manuelle de certaines données du profil permet d'avoir un minimum de connaissances sur l'utilisateur, puis le système emploie différents agents chargés de collecter de façon autonome les données des activités de l'utilisateur, par interaction avec les composants logiciels qu'il utilise (contenu du presse papier, page web visitée, document édité, temps passé sur chaque document, etc.). Certaines données du profil seront automatiquement déduites à partir l'environnement matériel et logiciel de recherche de l'utilisateur.

III.2.1.2 Le gestionnaire du profil document

Ce module intègre le processus d'indexation des documents en fonction des métadonnées telles que celles définies par le Dubin core. Les descripteurs obtenus de cette indexation représentent le profil de document. Le SRIP va manipuler deux structures représentant la sémantique du contenu des documents : le descripteur habituel issu de l'indexation classique et le profil du document obtenu par indexation par des métadonnées.

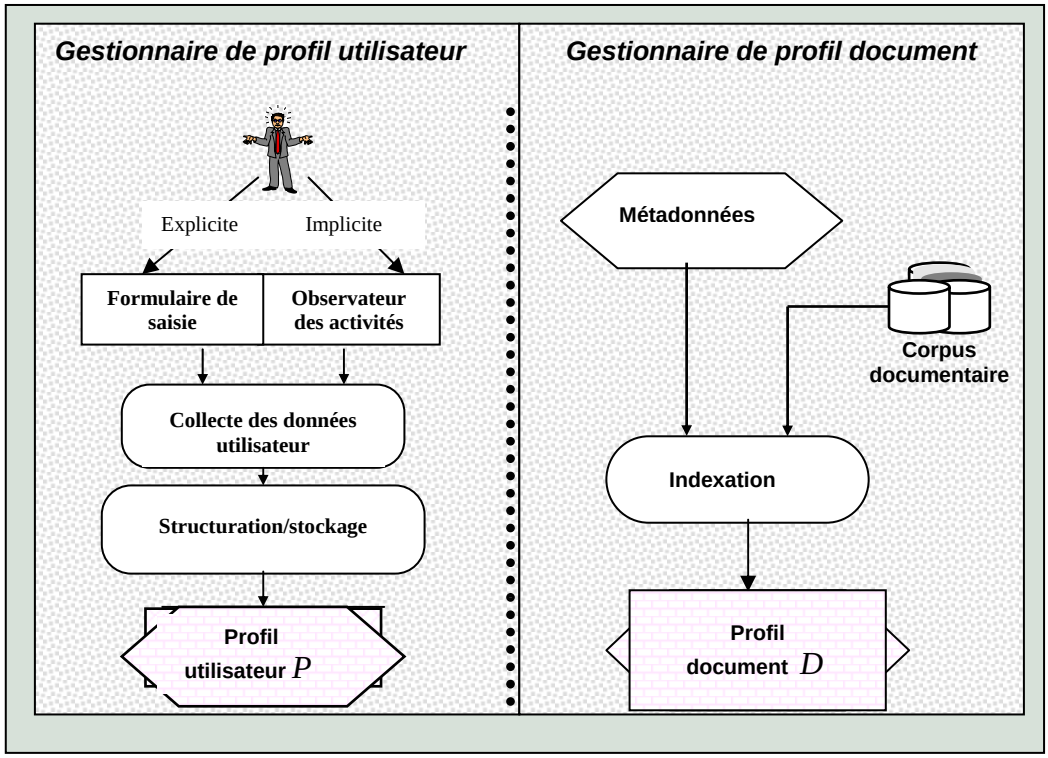

Figure 3.3 : *Module de construction de profils.*

III.2.2 Module d'intégration de profils

Ce module a pour rôle d'intégrer les profils dans le processus de recherche. Ce processus de recherche s'effectue en trois phases : l'analyse de la requête, la présélection des documents puis l'appariement entre la requête et les documents.

Ce module est donc décomposé en trois sous modules ayant chacun un rôle spécifique en fonction de la phase de recherche (Figure 3.4) :

- le gestionnaire de l'espace de recherche,
- le gestionnaire de la requête,
- le gestionnaire de l'appariement requête document.

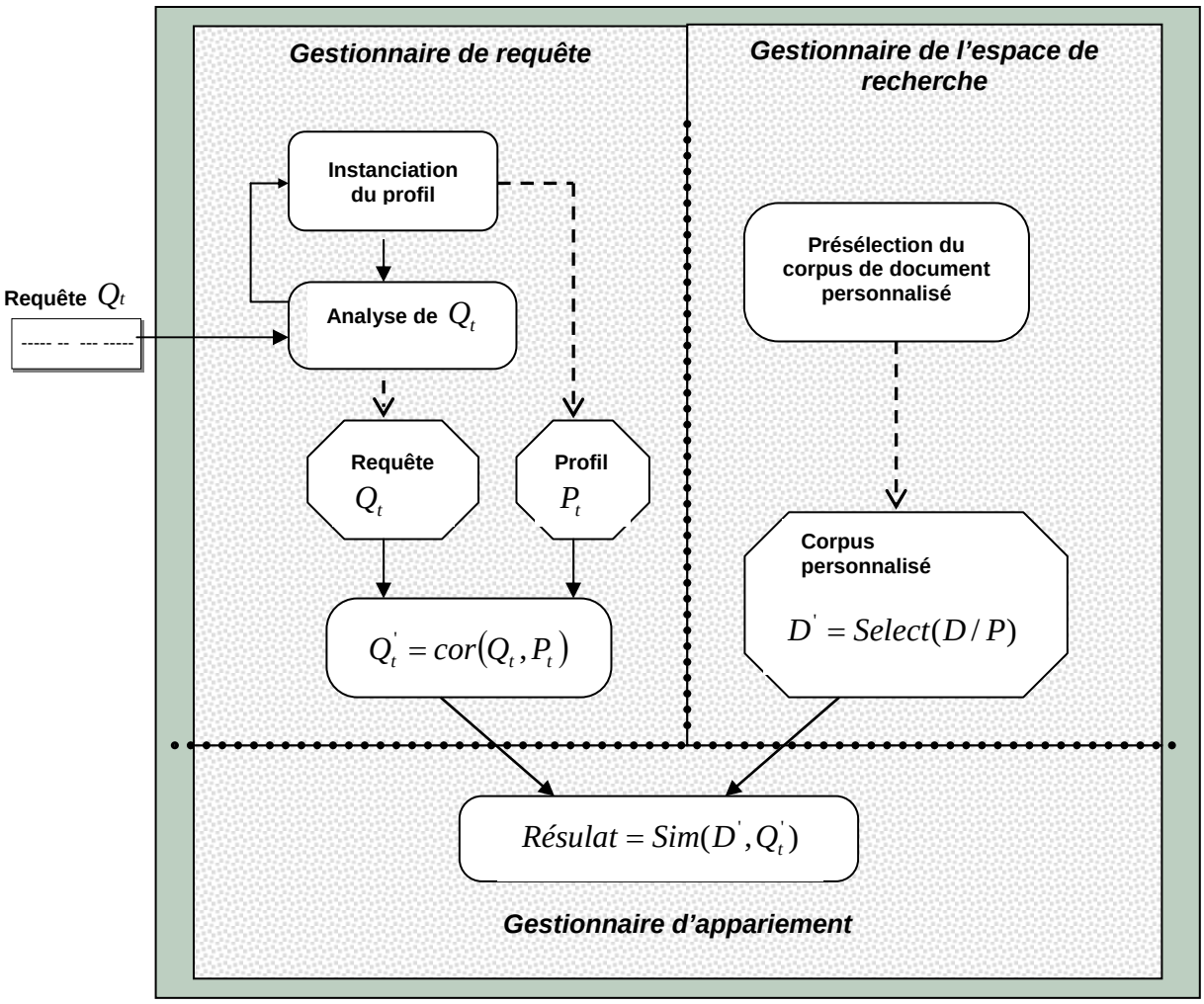

Figure 3.4 : *Module d'intégration de profils.*

III.2.2.1 Le gestionnaire de l'espace de recherche

La gestion de l'espace de recherche consiste à cibler dans l'ensemble du corpus le sous-ensemble de documents pertinents correspondant au profil de l'utilisateur. La description des documents par des métadonnées appariées avec le profil de l'utilisateur va permettre de présélectionner un corpus de documents pertinents, corpus sur lequel portera la requête.

Le gestionnaire doit donc définir une fonction de sélection des documents à partir des données du profil. On établit ainsi une corrélation sémantique entre le profil et les documents susceptibles d'être pertinents. Les préférences des utilisateurs sur l'origine de l'information seront prises en considération à ce niveau.

III.2.2.2 Le gestionnaire de la requête

Pour cibler le besoin réel de l'utilisateur, le système doit analyser la requête exprimée par l'utilisateur. De cette analyse, le gestionnaire de requête doit déduire les données nécessaires pour instancier le profil utilisateur.

Il est clair que si l'utilisateur formule sa requête initiale en tenant compte de ce qu'il est c'est-à-dire de son profil ; le système pourra mieux interpréter son besoin et donc mieux le satisfaire.
Partant de cette constatation, notre idée est d'essayer au maximum d'avoir une requête optimale dés le début de la recherche. On sous entend par optimale le fait de couvrir au mieux le contexte et le but de la recherche.
On pose comme heuristique que l'utilisateur, conscient de son rôle dans le processus de recherche, acceptera de fournir un minimum d'efforts pour formuler sa requête et qu'il y sera fortement guidé.
A partir de cette hypothèse, la requête initiale ne sera plus constituée d'une simple liste de mots clés que l'utilisateur formule, mais également des valeurs de propriétés du profil. Les valeurs de paramètres de certaines catégories du profil seront directement spécifiées par l'utilisateur. La requête ainsi améliorée par les données du profil est nommée « ***requête profilée*** ».
Le choix des paramètres est fortement lié au type d'application, cependant il est évident que les paramètres les plus significatifs du contexte de la recherche sont ceux liés aux catégories : « Domaine d'intérêts » et « Profession ».

Le gestionnaire doit donc inclure un mécanisme de correspondance entre requête utilisateur et profil utilisateur. Après avoir ciblé les données du profil qui englobe la recherche actuelle, le gestionnaire doit extraire ces données dans une forme compatible avec la requête.
Ces données vont servir à reformuler la requête initiale. On obtient donc une nouvelle requête plus proche du but de la recherche de l'utilisateur.

III.2.2.3 Le gestionnaire d'appariement

En exploitant le résultat des précédants gestionnaires à savoir : Un corpus documentaire personnalisé et une requête étendue en fonction du profil, le rôle de ce gestionnaire est d'effectuer l'appariement entre les deux. Le gestionnaire calcule ainsi une fonction de similarité entre le document, la requête et le profil utilisateur. Cette fonction correspond à une mesure de probabilité de pertinence d'un document sachant la requête et le profil utilisateur. Seuls les documents dépassant un seuil de similarité vont être sélectionnés.

III.2.3 Module de présentation du résultat

Une fois que l'information pertinente a été trouvée, il reste à la restituer à l'utilisateur. Le module de présentation du résultat est donc chargé de restituer l'information pertinente selon les préférences et les capacités de l'utilisateur (Figure 3.5).
Pour cela le module inclus des processus qui modifient l'aspect des documents et le moyen de restitution en fonction des données du profil.

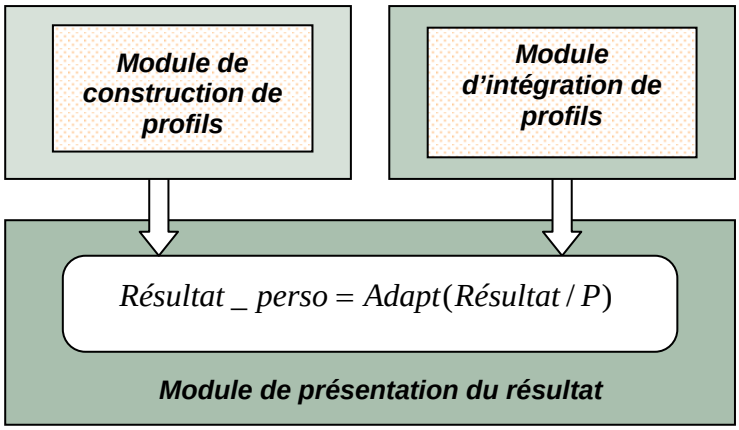

Figure 3.5 : *Module de présentation du résultat.*

III.2.4 Module d'évolution du profil

Le SRIP doit être capable de détecter les changements des centres d'intérêts et l'environnent de recherche de l'utilisateur en l'impliquant le moins possible (Figure 3.6).
C'est le module d'évolution du profil qui est chargé de détecter l'ensemble des informations issues soit des résultats des recherches précédentes de l'utilisateur soit, à partir d'un autre profil similaire, permettant de faire évoluer le profil de l'utilisateur.

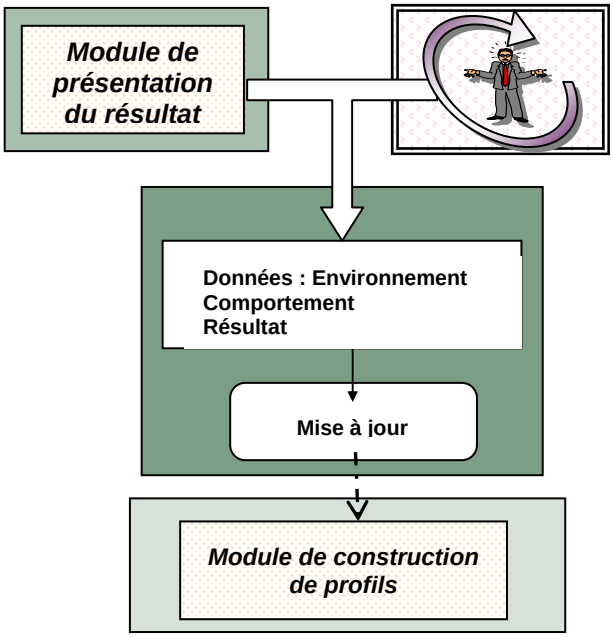

Figure 3.6: *Module d'évolution du profil.*

On récapitule dans la figure suivante (Figure 3.7) l'ensemble des modules composant le SRIP en mettant en évidence les relations existantes entre eux.

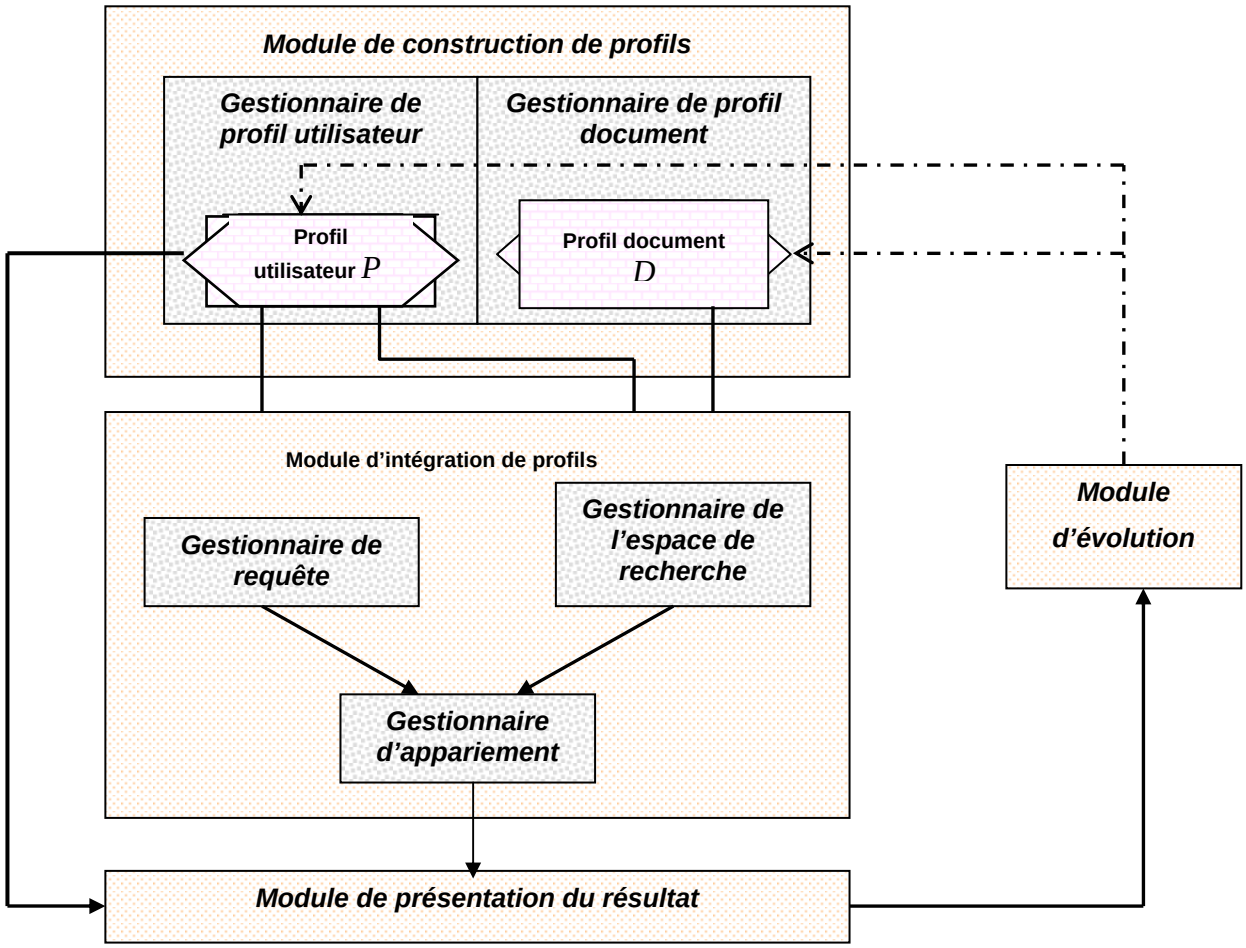

Figure 3.7 : *Architecture générale du SRIP.*

IV Interaction du profil utilisateur - processus du SRIP

Nous avons présenté dans les sections précédentes l'ensemble des catégories contenues dans le profil utilisateur ainsi que les différents composants du SRIP. Dans cette section nous abordons la phase d'intégration du profil utilisateur dans le processus d'accès personnalisé.

A cet effet, il faut tout d'abord déterminer quelles sont les catégories à exploiter et préciser dans quelles phases du processus.

Nous présentons dans ce qui suit le schéma global de correspondance entre profil utilisateur et architecture du SRIP (Figure 3.8).

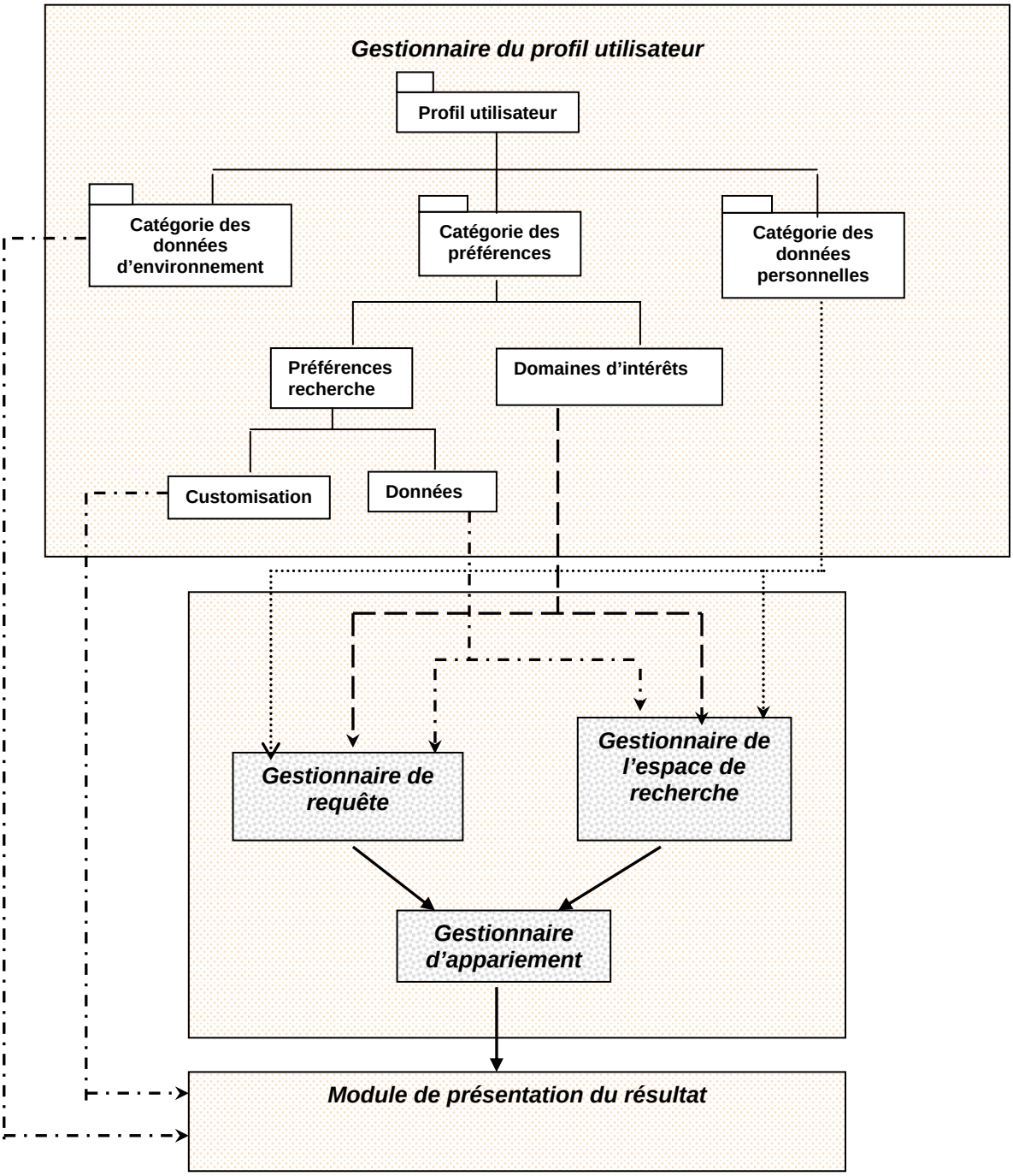

Figure 3.8 : Correspondance *profil utilisateur/module SRIP*.

A partir de ce schéma on peut très bien distinguer la relation existant entre les parties du profil et les composants du SRIP.

IV.1 Intégration du profil dans la phase de présélection de l'espace de recherche

Les catégories du profil utilisateur intégrées dans cette phase du processus sont :

- La sous catégorie « Domaine d'intérêts » qui contient au préalable l'ensemble des documents consultés par l'utilisateur et qui représente une base initiale de recherche.
- La sous catégorie « Contenu » qui contient les préférences de l'utilisateur concernant les types des corpus d'où il désire effectuer ou pas sa recherche.
- La sous catégorie « Profession » qui contient les informations sur le type de profession qu'exerce l'utilisateur.

A partir des informations contenues dans les sous catégories « Contenu » et « Procession » le SRIP va cibler un ensemble de documents. Ces informations sont représentées par une liste de termes, cette liste va être comparée aux descripteurs des documents de la collection. Les documents ayant un degré de similarité élevé seront sélectionnés et formeront le sous espace de recherche sur lequel portera la requête.

IV.2 Intégration du profil dans la phase d'évaluation de requête

La requête est l'expression d'un besoin circonstancié que l'utilisateur souhaite voir satisfait en tenant compte de son profil. Elle représente le besoin à court terme de l'utilisateur, alors que le profil représente son besoin à long terme.

Nous avons posé l'hypothèse que la requête initiale de l'utilisateur contient des valeurs de paramètres du profil utilisateur fournies directement par l'utilisateur. Ces paramètres représentent les classes d'intérêt de haut niveau dans la structure hiérarchique de la sous catégorie « Domaine d'intérêts ». En choisissant une classe d'intérêts l'utilisateur exprime son but de recherche. A partir des choix le système va extraire les documents correspondants et effectuer la reformulation de la requête. En effet, le SRIP va considérer ces documents comme jugés pertinents afin de reformuler la requête initiale.

IV.3 Intégration du profil dans la phase de présentation du résultat

Dans cette phase l'interaction entre le profil et le module de représentation du résultat se fait à travers les catégories « Customisation » et « Catégories des données de l'environnent ». En fonction des informations fournies par ces catégories le SRIP va déployer des mécanismes adaptation. La définition de ces mécanismes dépasse le cadre de notre travail.

V Conclusion

Nous avons abordé dans ce chapitre la problématique générale de l'accès personnalisé à l'information, en essayant de résoudre les problèmes liés à la représentation et l'intégration de utilisateur comme composante principale dans un processus de RI, nous avons abordé le problème de manière générale pour tenter d'approcher des solutions pour la mise en œuvre de tout le processus de recherche personnalisé. Une réflexion générale de ce que doit être un SRIP nous a conduit à définir les phases du processus du SRIP ainsi qu'une architecture pouvant supporter l'ensemble de composants de ce système. A partir des phases du processus de personnalisation on a pu proposer un profil utilisateur couvrant autant que possible l'ensemble des connaissances caractérisant l'utilisateur. On a donc catégorisé le contenu du profil en fonction du rôle de chaque catégorie dans le processus du SRIP. De cette intégration, on peut remarquer que certaines catégories telle que la catégorie des préférences a une incidence majeure sur les deux phases primordiales du processus à savoir la gestion de la requête et la gestion de l'espace de recherche.

La catégorie des préférences et la catégorie des données personnelles sont les plus importantes du profil. Les autres catégories ont une incidence minime dans le processus de recherche; elles n'opèrent qu'à la fin de la recherche de documents pertinents pour augmenter le degré de personnalisation.

CONCLUSION GENERALE

Les travaux présentés dans ce rapport s'inscrivent dans le cadre d'un projet de développement d'un SRIP permettant un accès personnalisé à l'information pertinente en fonction de chaque utilisateur. Le SRIP intégrera fortement la composante utilisateur tout au long de la chaîne d'accès à l'information pertinente.

Dans le cadre de notre travail on s'est le plus penché sur ce que peut représenter le profil de l'utilisateur et sur son impact dans l'ensemble du processus de recherche d'information. Il est clair que l'accès personnalisé à l'information passe par l'intégration de l'utilisateur dans toutes les étapes du chemin d'accès. Pour pouvoir intégrer l'utilisateur comme composante principale il faut avant tout définir des modèles et des structures manipulables par le système que l'on nomme « profil utilisateur ». Mais il faut également définir ce que va contenir ce profil comme informations pouvant influencer la recherche.

La mise en œuvre d'un SRIP permettant un accès personnalisé à l'information pertinente pour chaque utilisateur nécessite une large vision de ce qu'est la personnalisation.
Dans la première partie de ce rapport, nous avons présenté les grands axes de recherche menés par la communauté de RI pour aider l'utilisateur lors de sa recherche et pour adapter la recherche en fonction de son profil. Nous avons également présenté une synthèse regroupant les avantages et les inconvénients de chacune des approches. Il en ressort de cette synthèse que l'on s'oriente de plus en plus vers des systèmes de personnalisation.

Une étude des systèmes personnalisés existants nous a permis d'effectuer une synthèse de ce que représente un système de recherche personnalisé. L'ensemble des concepts clés d'un SRIP ainsi que les approches de mise en œuvre sont présentés dans la deuxième partie de ce rapport. La difficulté de ce travail se situe dans la diversité des techniques de personnalisation employées par chaque système et le fait de faire ressortir de ces systèmes les concepts clés communs pour tout système de recherche personnalisé.

En conclusion des deux premières parties, il apparaît clairement que la personnalisation de l'information a été abordée depuis longtemps mais de façon parcellaire dans différents types d'applications chacune apportant une définition du profil utilisateur.
Dans la dernière partie, nous avons abordé la problématique générale de l'accès personnalisé à l'information, en essayant de résoudre les problèmes liés à la représentation et l'intégration de l'utilisateur comme composante principale dans un processus de RI. Une réflexion générale de ce que doit être un SRIP nous a conduit à définir les phases du processus du SRIP ainsi qu'une architecture pouvant supporter l'ensemble des composants de ce système.

En conclusion, ce travail apporte une définition claire de ce que doit être un système de recherche d'information personnalisé et comment doit s'effectuer cette personnalisation indépendamment de tous types de systèmes. La représentation du profil utilisateur sous la forme d'une hiérarchie de catégories permet de regrouper l'ensemble des connaissances caractérisant l'utilisateur nécessaire à la personnalisation de sa recherche. Cette catégorisation du contenu du profil est faite en fonction du rôle de ce contenu dans chaque phase du processus de personnalisation.

En perspective à ce travail, nous envisageons :
- De proposer un modèle formel pour représenter le profil utilisateur. Ainsi que la détermination effective des paramètres du modèle.
- De spécifier les mécanismes de construction et de mise à jour du profil ainsi que l'ensemble des fonctions de présélection de l'espace de recherche en fonction du profil, de reformulation de requête et d'appariement entre requête document et profil.
- De mettre en place une plate forme d'implémentation de l'architecture du système proposé. On s'intéressera plus particulièrement aux phases de la présélection de l'espace de recherche, de l'évaluation de la requête et d'appariement qui combineront les trois paramètres : document, requête et profil utilisateur.
- De développer un prototype d'évaluation de l'impacte de l'intégration de ce profil dans le processus d'accès personnalisé à l'information et sur les performances de recherche d'un tel système.

BIBLIOGRAPHIE

[Amato 99] G. Amato, U. Straccia, *User Profile Modeling and Applications to Digital Libraries*, Proc. 3rd European Conf. Research and Advanced Technology for Digital Libraries, ECDL, 1999.

[Allen 90] R. Allen, User models, Theory, methods and practice. INT. J. Man-Machine Stud. 1990.

[Amati 97] G. Amati, F. Crestani, F. Ubaldini, *A learning system, for selective dissemination of information*. Proceeding of I JCAI'97, 1997.

[Bradford 99] C. Bradford and I. Marshall. *Analysing Users WWW Search Behaviour*. Proc IEE colloquium *99/149* - navigating the web.1999.

[Ben Abdallah 97] N. Ben Abdallah, *Analyse et structuration de documents scientifiques pour un accès personnalisé à l'information : Vers un système d'information évolué.* Thèse de doctorat, Université Lyon 1, 1997.

[Belkin 92] N.J, Belkin, W.B. Croft, *Information Filtering and Information Retrieval: Two sides of the Same Coin?* Communication of the ACM, 35(12), pp 29-38, 1992.

[Bloedorn 96] E.Bloedorn, I. Mani, and T.R. Macmillan, *Machine Learning of User Profiles: Representational Issues*. Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96), Portland, OR, AAAI/MIT Press, August, 1996 .

[Bloomberg 02] Bloomberg, Financial Services, http://www.bloomberg.com.

[Bottraud 04] J.C. Bottraud , G. Bisson , M.F. Bruandet, *Apprentissage de profils pour un agent de recherche d'information*. Coria'04, IRIT Toulouse France, 2004.

[Boughanem 98] M .Boughanem, C. Chrisment, C. SOULE-DUPUY, *Query modification based on Relevance Back-propagation in ad-hoc environment*. IPM: Information Process and Mangement 1998.

[Boughanem 04] M. Boughanem, H. Tebri, M. Tmar. *Filtrage d'information*. Dans : *Méthodes avancées pour la recherche d'informations*. Madjid Ihadjadene (Eds.), Hermes-Lavoisier, 11, Rue Lavoisier 75008 Paris, hermes science, 137-162, 2004.

[Bourne 79] C.Bourne & B.Anderson, *DIALOG LabWorkbook*. Second edition, Looked Information Systems, PaloAlto, Californie, USA. 1979.

[Chen 01] Chien Chin Chen, Meng Chang Chen, Yeali *Sun: A Web Document Personalization User Model and System Machine Learning*, Information Retrieval and User Modeling - Workshop Schedule UM-2001.

[Claypool 01] M. Claypool, L. E. Phong, M. Wased, D. Brown, *Implicit Interest Indicators*. IUI'01, Santa-Fe, New Mexico, USA, 2001

[Croft 93] W.B. Croft , *Knowledge-based and Statistical approaches to Text Retrieval*. IEEE EXPERT, vol. 8, n° 2, Avril, 1993.

[Croft 83] W. B. Croft, *Experiments with representation in a document retrieval system*. Information Technology : Resaerch and Development, Vol 2, N° 1, 1983.

[Cnn 04] CNN, News Channel, http://www.mycnn.com.

[Declaris 94] N. DeClaris, J. James, A. Nerode, W. Kohn, Intelligent integration of medical models, Proc. IEEE Conference on Systems, Man, and Cybernetics, San Antonio,1994.

[Denning 82] P.J.Denning, *Electronic Junk*. Communication of the ACM, Vol 25, N°3, 1982.

[Dublin 04] http://purl.org/metadata/dublin_core.

[Deitel 01] H.M. Deitel, P.J. Deitel, K. Steinbuhler, *e-Business and e-Commerce for Managers*. Prentice-Hall, 2001.

[Dan 86] J.P Daniels , *Cognitive Models in Information Retrieval – An Evaluation Review*. Journal of documentation, vol 42, n° 4, December 1986, p 272-304.

[Neuhold 03] J. E. Neuhold, *Personalization and User profiling & Recommender Systems*. WI/IM, Information management Proseminar 2003.

[Fir 04] FireFly, http://www.firefly.net. 2004.

[Fluhr & al, 1985] C.Fluhr & F.Debili, *Interrogation en Langue Naturelle de Données Textuelles et Factuelles*. Intelligent Multimédia Information System and Management (RIAO), Grenoble France, 1985.

[Inf 04] Info Quest, http://www. Infoquest.com 2004.

[Gauch 03] S. Gauch, J. Chaffe, A. Pretschner, *Ontology-Based User Profiles for Search and Browsing, To appear in J. User Modeling and User-Adapted Interaction*, the Journal of Personalization Research , Special Issue on User Modeling for Web and Hypermedia Information Retrieval, 2003

[Gessler 93] N. Gessler , *Gearge Boole et l'algèbre de la logique*. Etudes logiques, Neuchâtel,1993.

[Grossman 93] D. A. Grossman, O. Frieder, *Information retriveval, Algorithms and heuristics.* Kluwer Academic Publishers 1993.

[Goldberg 00] K.Goldberg, T. Roeder, D. Huptan, C. Perkins. Eigentaste , *A constant time collaborative filtering algorithm*. Technical Report M00/41, IEOR and EECS Departements, UC Berkeley, Août 2000.

[Goecks 00] J. Goecks and J. Shavlik. *Learning Users' Interests by Unobtrusively Observing Their Normal Behaviour*. In Proceedings of the 2000 International Conference on Intelligent User Interfaces, New Orleans, LA, 2000.

[Grossman et Frieder, 1998] D. Grossman, O. *Frieder: Information Retrieval. Algorithms and Heuristics*. Kluwer Academic Publishers, 1998.

[Haman 92] D. Harman, *Relevance feedback revisited*. Proceedings of ACM SIGIR1992.

[Huang 98] M. Huang , *Extracting Classification Knowledge of Internet Documents with Mining Term Associations*. A semantic Approach. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 1998.

[Kurki 99] T. Kurki, S. Jokela, R. Sulonen, M. Tirpeinen, Agents in delivering personalized content based on semantic metadata, In Proc, AAAI S*ping Symposium*, 1999.

[Kruliwich 97] B. Kurulwich, C. Burkey, *The Info Finder Agent, Learning User Interests through Heuristic Phrase Extraction*, IEEE Expert, September/October 1997.
[Lelu 92] A Lelu, C. François : *Automatic generation of hypertext links in information retrieval systems*. Communication of colloque ECHT'92 , ACM Press, New York, 1992.

[Luhn 57] H. P. Luhn, *A statistical approache to mechanized encoding and searching of literary information*. IBM, vol. 1, N° 4, 1957.

[Huhn 99] M. N. Huhn, L.M. Stephens, *Personal Ontologies*. Internet Computing, Vol. 3, No. 5, pp. October 1999

[Miller 97] B.Miller, J. Konstan, D. Matlz, J.L. Herlocker, L. Gordan, A. Riedl, *GroupLens : applying callaborative filtring Usenet news*. Communications of ACM, March 1997.

[Mothe 92] J. Mothe: *Recherche et exploration d'information découverte de connaissances pour l'accès à l'information.* Thèse de l'Université Paul Sabatier; 1992.

[Pazzani 96] M. Pazzani, J. Muramatsu, D. Billsus, Syskill, *Webert: Identifying Interesting Web Sites*. In Proceedings of the 13th National Conference On Artificial Intelligence, 1996.

[Morita 94] M. Morita, Y. Shinoda, *Information filtring based on user behavior and best match text retrieval*. Proceeding of ACM SIGIR 1994.

[Perenon 00] Pascal Perenon, Mémoire de D.E.A, *Réalisation d'un prototype de Système de Recherche d'Informations Scientifiques* Laboratoire RECODOC, Université Claude Bernard Lyon 1 Juillet 2000.

[Pednault, 2000] P.D. Pednault, *Representation is Everything*. Communications of the ACM, August 2000.

[Poi 04] Point Cast http://www.pointcast.com.

[Pretschner 99b] Alexander Pretschner, Susan Gauch. *Ontology Based Personalized Search*. In Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), November 1999.

[Poo 03] Danny Poo, Brian Chng, Jie-Mein Goh, *A Hybrid Approach for User Profiling*. 36th Annual Hawaii International Conference on System Sciences (HICSS'03) - Track 4, January 06 - 09, 2003.

[Robertson 76] S.Robertson & K.Sparck Jones, *Relevance Weighting for Search Terms* . Journal of The American Society for Information Science, Vol 27, N°3, 1976

[Robertson 99] S.E.Robertson, S.Walker, M.Beaulieu, *Automatic Adhoc Filtring, VLC and Interactive Track*. In Poceeding of the 7th Text Retrieval Conference TREC7, 1999.

[Robertson 00] S. Robertson, S. Walker. *OKAPI/Keenbow at TREC-8* 2000.

[Rocchio 71] J. Rocchio : *Relevance feedback information retrieval*. In Gerald Salton (editor), The SMART retrieval system- experiments in automated document processing. Prentice-Hall, Englewood Cliffs, NJ, 1971.

[Search 02] Search Engine Showdown Size statistics. http://www.searchengineshowdown.com/ stats/sizeest.shtml.

[SALT, 71] G. Salton. *The SMART Retrieval System: Experiments in Automatic Document Processing*, Prentice-Hall Inc, NJ.1971.

[SALT 89] G.Salton, *Automatic Text Processing*, The Transformation Analysis and Retrieval of Information by Computer. 1989. Addison Wesley.

[SALT 90] G. Salton & C. Buckley. *Improving Retrieval Performance By Relevance Feedback*, Journal of The American Society for Information Science. 1990.

[Salton 94] G. Salton & J. Allan, *Automatic Text Decomposition and Structuring*. Actes du Congrés RIAO'94, Intelligent Multimedia Information on Retrieval Systems and Management, New York. 1994 .

[Singhal 97] A. K. Singhal, *Term weighting revisited*, PHD of Cornell University 1997.

[Shavlik 99] 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, Redondo Beach, CA, 1999.

[Somlo 03] G L. Somlo, A. E. Howe, *Using Web Helper Agent Profiles in Query Generation International Conference on Autonomous Agents.* Proceedings of the second international joint conference on Autonomous agents and multi agent systems Melbourne, Australia Web technologies. 2003

[Sparch Jones 79] K. Sparck Jones *Experiments in relevant weighting of search terms*. IPM 1979.

[Tmar 02] M. Tmar. Modèle auto-adaptatif de filtrage d'information : apprentissage incrémental du profil et de la fonction de décision. Thèse de l'Université Paul Sabatier. Toulouse 2002.

[TREC9, 2000] proceedings of the Ninth Text R Etrieval Conference (TREC-9) held in Gaithersburg, Maryland, November13-16, 2000.
URL : *http://trec.nist.gov/pubs/trec9/t9_proceedings.html*.

[Vassiliou 02] CH. Vassiliou, D. Stamoulis, D. Martakos, *The process of personalizing web content: techniques, workflow and evaluation*. Proceedings of the SSGRR International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet. 2002.

[Walker 97] S. Waller, S. E. Robertson, M. Boughanem, G. J. F. Jones, K. Sparck Jones. *Okapi at TREC-6 automatic and ad hoc,* VLC routing, filtring and QSDR. Proceeding of TREC-6, 1997.

[Yates 99] R. B.Yates, R. Neto, *Modern Inforation Retrival*. ACM Press, Addison Wesley, 1999.

[Yho 04] Yahoo! http://www.yahoo.com.2004.