REPUBLIQUE *ALGERIENNE DEMOCRATIQUE ET POPULAIRE*

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Mouloud MAMMERI de Tizi-Ouzou

Faculté de Génie Electrique et Informatique
Département D'Informatique

Mémoire de fin d'études

En vue de l'obtention du diplôme de Master en Informatique

Spécialité : ingénierie des systèmes d'information

Thème

Evaluation de l'impact d'intégration du profil utilisateur dans la recherche d'information par l'utilisation des profils des utilisateurs réels.

Encadré par:

M^me^ ACHEMOUKH Farida

Réalisé par :
- ABKARI Tassadit

- REHOUNE Lynda

Mémoire soutenu, devant le jury composé de :

Présidente: *M^me^ G.SINI.*
Promotrice: *M^me^ F.ACHEMOUKH.*
Examinatrice : *M^me^ S.FELLAG.*

Remerciement

Nous tenons tout d'abord à remercier Dieu le tout puissant et miséricordieux, qui nous a donné la force et la patience d'accomplir ce modeste travail.

En second lieu, nous tenons à remercier notre Promotrice : Madame ACHEMOUKH, ses précieux conseils Et son aide durant toute la période du travail.

Nos vifs remerciements vont également aux membres du jury pour l'intérêt qu'ils ont porté à notre recherche en acceptant d'examiner notre travail et de l'enrichir par leur propositions.

Nous aimerions également exprimer notre gratitude à tous ceux qui ont contribué de prés ou de loin pour ce Projet soit possible,

On vous dit merci

Dédicace

Je dédie ce modeste travail à

A celle qui a bercé mes rêves ma mère.

A celui qui a nourri mes ambitions mon père.

A celles qui a soulevé bien des fardeaux avec moi ma sœur ATIKA.

A mes anges gardiens mes frères Amine, Remdane et Samy.

A mes chers Hamida et Massinissa pour leur encouragement.

A mon binôme et sa famille.

A tous ce qui me connaissent.

Tous mes enseignants;

A toute la promotion 2018/2019.

Tous ceux et toutes celles qui m'ont soutenue de près ou de loin tout au long de mon parcours universitaire.

LYNDA

Dédicaces

Je dédie ce modeste travail à :

Mes chers parents qui m'ont soutenue, entourée et motivée pour avancer dans la vie et devenir celle que je suis aujourd'hui;

Ma grand-mère;

Mes frères et sœurs qui ont toujours été à mes côtés et m'ont assistée dans les moments difficiles;

Toute ma famille et tous mes amis;

Mon binôme Lynda et sa famille;

Tous mes enseignants;

Tous ceux et toutes celles qui m'ont soutenue de près ou de loin tout au long de mon parcours universitaire.

TASSADIT

Liste des figures

Liste des tableaux

Sommaire

Chapitre 03 : Evaluation des systèmes d'accès personnalisé à l'information

Chapitre 04 : Modélisation de la personnalisation à base de profils réels.

Résumé

La surabondance de l'information ainsi que sa large accessibilité à travers le web a engendré une dégradation des performances des systèmes de recherche d'informations (SRI).

L'origine de ce problème réside en partie dans le caractère non personnalisé du processus d'accès à l'information. La personnalisation et une dimension qui permet la mise en œuvre de système centrée utilisateur, en vue d'adapté son fonctionnement a son contexte précis, ses préférences ou plus globalement, son profil.

Ce sujet se situe précisément dans le contexte de l'accès personnalisé à l'information.

Tel que des utilisateurs ont des objectifs différents, des contextes différents et perceptions différentes de la notion de pertinence.

Notre objectif est d'évaluer l'impact d'intégration du profil utilisateur dans la recherche d'informations par l'utilisation des profils des utilisateurs réels.

La démarche méthodologique sera articulée sur :

1- Étude exploratoire sur les facteurs intervenant dans la définition d'un profil utilisateur dans un SRI.
2- Évaluation expérimentale du modèle de profil proposé. En utilisant une collection web via une interface (tel que Google) des vrais utilisateurs qui formulent leurs requêtes selon un besoin spécifique. Les interactions des utilisateurs tels que les clics, temps passé sur une page, etc., sont enregistrées et exploitées dans l'étape d'évaluation de performance du système.
Les documents cliqués sont utilisés afin de construire le profil utilisateur.

Mots-clés :

Recherche d'informations, Recherche d'informations personnalisées, Modélisation d'un utilisateur, Centre d'intérêt, Évaluation d'un système de recherche d'informations.

INTRODUCTION GENERALE

Introduction Générale

La recherche d'information (RI), est une branche en informatique qui s'intéresse à l'acquisition, l'organisation, le stockage et la recherche des informations. C'est l'ensemble de procédures et techniques permettant de sélectionner, parmi un ensemble de documents, les informations (documents ou parties de documents) pertinentes en réponse à un besoin en information exprimé par l'utilisateur à travers une requête.

Les premier travaux dans le domaine de la RI se sont focalisée à résoudre des problèmes principalement liées à la représentation de l'information, l'évaluation des requêtes ainsi que l'évaluation de performance de recherche.

Toutefois, le processus d'appariement, la différence du vocabulaire utilisé pour l'expression du contenu des documents et des besoins en information n'est pas prise en compte.

Ce défaut d'appariement engendre la dégradation de performance de recherche. Ainsi le problème n'est plus étant la disponibilité de l'information mais la capacité de sélectionnée l'information répondant aux besoins d'un utilisateur.

Afin de résoudre ce problème, les t ravaux s'orientent vers la RI personnalisée qui consiste à la révision de la chaine d'accès à l'information dans la perspective d'intégrer l'utilisateur dans l'ensemble des phases de recherche, afin de lui délivrer l'information pertinente adaptée à son contexte et ces préférences répondant à ces besoins précis ou plus globalement, son profil.

Dans le cadre de ce mémoire nous nous intéressons à l'impact d'intégration du profil utilisateur dans la recherche d'information par l'utilisation des profils des utilisateurs réels.

La définition d'une méthodologie d'évaluation reste une problématique majeure dans le domaine de la recherche d'information personnalisée, aucune collection de test standard n'a été construite pour évaluer l'efficacité de l'accès personnalisée à l'information.

En plus de l'absence de collection de tests, la recherche dans ce domaine est confrontée à l'inexistence de méthodologies, de mesures standards d'évaluation de l'adéquation des profils appris aux centres d'intérêts de l'utilisateur, ni l'existence de système référentiel.

Il est d'autant plus difficile de réalisé des scénarios d'évaluation objectifs en intégrant la dimension de l'utilisateur dans le processus d'accès.

Notre travail s'inscrit dans la perspective de créé une collection de test, à partir de laquelle on va définir des profils utilisateurs dans le but de vérifier leur impact sur la recherche d'information.

Ce mémoire est organisé en quatre chapitres :

- Dans le premier chapitre, nous présentant le passage de la RI classique vers la RI personnalisé en passant par la RI adaptive.
- Le second chapitre traite principalement la recherche d'information personnalisée, modélisation de l'utilisateur, intégration de profil utilisateur dans les différentes phases de processus globale d'accès personnalisée à l'information.
- Le troisième chapitre présente la problématique liée à la mise en place d'une compagne d'évaluation pour l'accès personnalisé ainsi qu'une synthèse des approches d'évaluation utilisées en RI personnalisée.
- Dans le quatrième chapitre nous présentons notre approche qui englobe nos contributions pour la création d'une collection de test réel et la définition des profils utilisateurs à partir de cette collection et nous discuterons les résultats de l'impact de l'intégration de ces profils dans le processus de recherche d'information.

Chapitre 1

De la RI classique à la RI personnalisée

1.1 Introduction

Le domaine de la recherche d'information est lié à la représentation de l'information, l'évaluation de requêtes ainsi que l'évaluation des performances de recherche. Le processus d'appariement dans les systèmes liés à ce domaine ne prend pas en compte le besoin en information réel de l'utilisateur.

Ce défaut d'appariement engendre une dégradation des performances de recherche. Ainsi, le problème qui se pose actuellement n'est plus tant la disponibilité de l'information mais la capacité d'accès et de sélection de l'information.

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

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

Afin de mieux cerner cette évolution, nous discutons dans ce chapitre des principaux facteurs et problématiques ayants conduits à l'émergence de la RI personnalisée. Plus particulièrement, nous abordons les problématiques majeures de la RI classique, ensuite nous présentons l'orientation des travaux vers la RI adaptative pour laquelle nous donnons un aperçu des techniques développées ainsi que les limitations inhérentes qui ont mené à la RI personnalisée.

1.2 Les fondements de la recherche d'information

Un système de recherche d'information (RI) est un système qui permet de retrouver, à partir d'une collection de documents, les documents susceptibles d'être pertinents à un besoin en information d'un utilisateur exprimé sous forme d'une requête. Dans cette définition, il y a trois notions clés que nous allons présenter dans la suite : document, requête, pertinence. (Daoud & al , 2009)

1.2.1 Notions de base

– Document

On appelle document toute unité d'information qui peut constituer une réponse à un besoin en information/requête d'un utilisateur. Un document peut être un texte, un morceau de texte, une image, une bande vidéo, etc.

– Requête

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

On peut citer :

- par une liste de mots clés
- en langage naturel
- en langage booléen.
- en langage graphique.

– Pertinence

La pertinence est la correspondance entre un document et une requête ; une mesure de l'informativité du document à la requête, un degré de relation entre le document et la requête …etc.

1.2.2 Principales phases du processus de RI

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

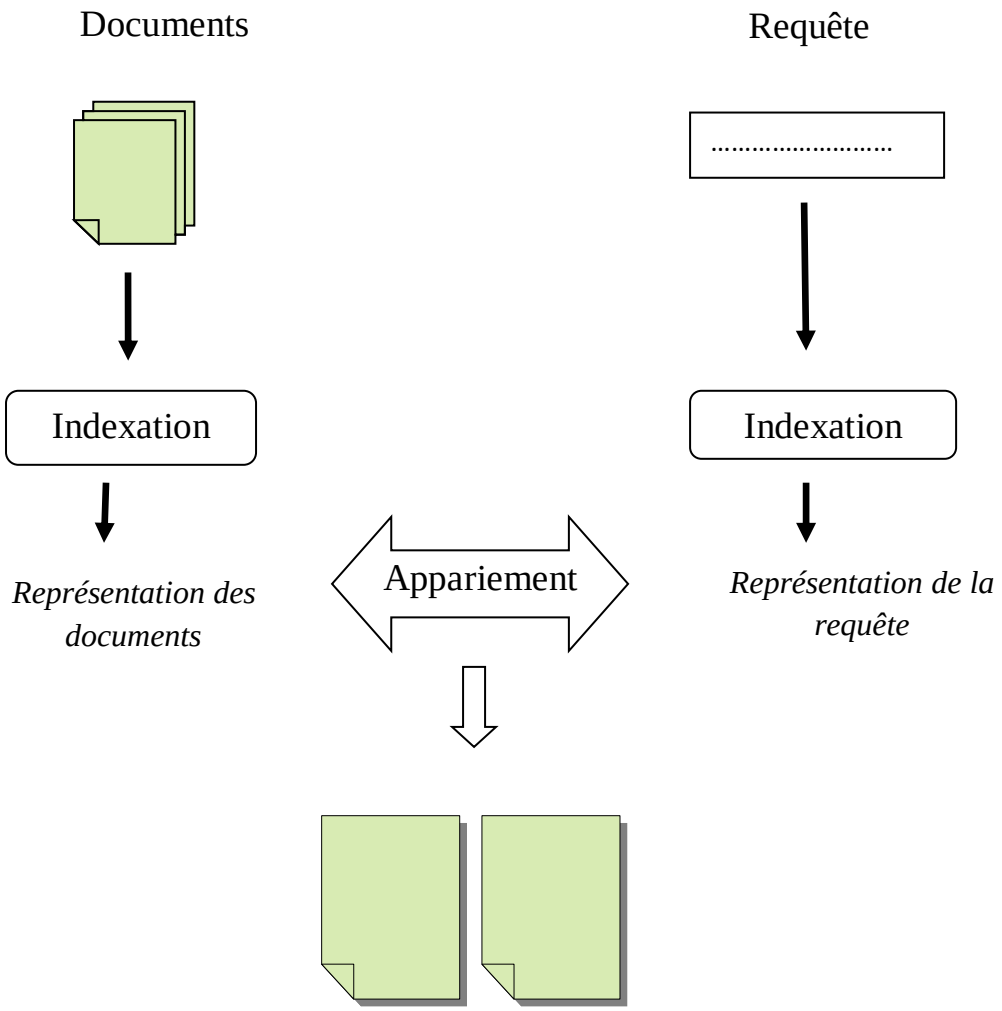

Figure 1.1. Processus en U de la RI (N.Zemirli & al, 2008)

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

1.2.2.1 L'indexation

L'indexation est une étape très importante dans le processus de RI. Elle consiste à déterminer et à extraire les termes représentatifs du contenu d'un document ou d'une requête. La qualité de la recherche dépend en grande partie de la qualité de l'indexation. Le résultat de l'indexation constitue, ce que l'on nomme le **descripteur** du document ou de la requête. Ce dernier est souvent une liste de termes ou groupe de termes significatifs pour l'unité textuelle correspondante, généralement assortis de poids représentant leur degré de représentativité du

contenu sémantique de l'unité qu'ils décrivent. Les descripteurs des documents (mots, groupe de mots) sont rangés dans un catalogue appelée dictionnaire constituant le **langage d'indexation**.

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

- ✓ ***Manuelle*** **:** chaque document est analysé par un spécialiste du domaine ou un documentaliste.

- ✓ ***automatique*** **:** chaque document est analysé à l'aide d'un processus entièrement automatisé.

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

Les termes extraits des documents ne jouent pas le même rôle dans la représentation de ces derniers, en ce sens où ils n'ont pas le même degré d'importance. Pour caractériser ce degré de discrimination, il est courant en RI, d'affecter à chaque terme un poids. Cette étape est primordiale dans le processus d'indexation correspond au processus pondération. Pour trouver les termes du document qui représentent le mieux son contenu sémantique.
La fonction de pondération d'un terme dans un document est connue sous la forme de *TF*IDF*. (Robertson & Jones, 1976)

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 :

TF **:** utilisation brute ou **log (1+*TF*).**

Où TF est la fréquence du terme dans le document.

- ***IDF* (*Inverse Document Frequency*) :** mesure l'importance d'un terme dans toute la collection. L'idée sous-jacente est que les termes qui apparaissent dans peu de documents de la collection sont plus représentatifs du contenu de ces documents que ceux qui apparaissent dans tous les documents de la collection.

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

$$\boldsymbol{IDF = \log\left(\frac{N}{df}\right)}, \quad \boldsymbol{IDF = \log\left(\frac{N-df}{df}\right).}$$

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 :

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

Une normalisation de la mesure du *TF _ IDF* par rapport à la longueur des documents a été proposée. Une des formules les plus utilisées (citées) aujourd'hui dans le domaine de la RI est la formule *BM*25 d'OKAPI tel que le poids d'un terme *i* dans le document *j* (noté *w* (*i*; *j*)) est donnée par :

$$\boldsymbol{w(i,j) = 0.5 * \frac{tfij * \log(\frac{N-ni+0.5}{ni+0.5})}{2*\left(0.25+\frac{0.75*dlj}{avgdl}\right)+tfij}} \qquad \textbf{(1.2)}$$

Où :

***ni* :** le nombre de documents contenant *ti*,

***N* :** le nombre de documents pertinents dans la collection,

***dl* :** la longueur du document *d j*,

***avgdl* :** la longueur moyenne des documents de la collection,

***tfi j* :** la fréquence d'apparition du terme *ti* dans le document *dj*.

1.2.2.2 L'appariement document-requête

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

***-Appariement exact* (« *exact match retrieval* »)**

Le résultat est une liste de documents respectant exactement la requête spécifiée avec des critères précis. Les documents retournés ne sont pas triés. (Salton, 1971)

***-Appariement approché* (« *best match retrieval* »).**

Le résultat est une liste de documents censés être pertinents pour la requête. Les documents retournés sont triés selon leur score de pertinence vis-à-vis de la requête. (Robertson & Jones, 1976)

1.2.3 Les modéles de RI

Les travaux de recherche dans le domaine de la RI ont conduit à la proposition de nombreux modèles. Un modèle de RI a pour rôle de fournir une formalisation du processus de recherche d'information et un cadre théorique pour la modélisation de la mesure de pertinence.

Nous présentons très brièvement dans ce qui suit les plus importants des modèles proposés.

Modèle booléen.

Ce modèle propose la représentation d'une requête sous forme d'une expression logique. Les termes d'indexation sont reliés par les connecteurs logiques *ET*, *OU* et *NON*. Le processus de recherche mis en œuvre, consiste à effectuer des opérations sur les ensembles de documents définis par la présence et l'absence de termes d'indexation, afin de réaliser un appariement exact avec l'équation de la requête. Une extension de ce modèle a été effectuée par (Salton & McGill., 1983): le modèle booléen étendu. Il intègre des poids dans l'expression de la requête et des documents. La sélection des documents s'effectuera donc sur la base d'un appariement rapproché et non plus exact.

Modèle vectoriel (Vector Space Model).

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

Modèle probabiliste (Probabilistic Model).

Ce modèle aborde le problème de la recherche d'information dans un cadre probabiliste. La pertinence document-requête est traduite par le calcul de la probabilité de pertinence d'un document par rapport à une requête. La pertinence entre un document et une requête est mesurée par le rapport entre la probabilité qu'un document *D* donné soit pertinent pour une requête *Q*, notée $p(R=D)$, et la probabilité qu'il soit non pertinent, notée $p(\bar{R}=D)$, ou R est l'élément de pertinence et $\bar{R}$ de non pertinence. Ces probabilités sont estimées par les probabilités conditionnelles selon qu'un terme de la requête est présent, dans un document pertinent ou dans un document non pertinent.

1.3 De la RI classique à la RI adaptative

La recherche d'information classique se base principalement sur le calcul de la pertinence du document selon des critères de sélection par le contenu et de la disponibilité de l'information ou alors elle peut également exploiter la structure des liens entre les documents afin de retourner une liste de résultats en réponse à une requête utilisateur. La limite majeure de la RI classique est qu'elle est basée sur une approche généraliste qui évalue invariablement les requêtes des utilisateurs et délivrent des résultats sans tenir compte des critères spécifiques de l'utilisateur qui a émis la requête ou du contexte de recherche. Par exemple, la requête "'gouvernement"' est une requête locale qui doit renvoyer des résultats concernant le gouvernement du pays dans lequel est localisé l'utilisateur. Par exemple, cette requête doit renvoyer en *France* le site du gouvernement français. La requête "*Java magazines*" est une requête ambigüe du fait qu'elle est liée à différentes thématiques de recherche, telles que des le langage de programmation *Java*, l'île de *Java* ou un magazine ayant le nom *Java*. Pour ce

type de requêtes, le système doit renvoyer des résultats dépendant des centres d'intérêts de l'utilisateur lors de la recherche.

L'accès à une information pertinente adaptée à des critères spécifiques de l'utilisateur et au contexte de recherche fait l'objet de la RI adaptative. Nous abordons dans la suite de cette section les facteurs d'émergence de la RI adaptative, les techniques développées en RI adaptative ainsi que leurs limitations de couvrir la notion large du contexte.

1.3.1 Facteurs d'émergence de la RI adaptative

Les facteurs d'émergence de la RI adaptative sont principalement lies a la prolifération des ressources d'information hétérogènes, la diversité et l'ambiguïté des besoins en information des utilisateurs ainsi que l'apparition de la RI mobile basée sur l'adaptation du processus de RI au contexte géographique de l'utilisateur.

A. Volume de l'information

L'accès à une information pertinente dans un environnement de recherche qualifie par la prolifération des ressources hétérogènes (données structurées, documents textuels, composants logiciels, images) est un défi réel pour la RI classique.

Le volume de l'information disponible sur le *Web* ainsi que le nombre d'utilisateurs sont toujours en croissance vertigineuse. En outre, les utilisateurs n'utilisent habituellement que quelques mots (environ 3 d'après (Research & al, 2001), (Jansen & al, 1998)) pour décrire le document recherché. Face au volume de l'information, la problématique majeure des SRI classiques est qu'ils proposent en réponse une liste massive de documents, ayant tous des estimations de pertinence comparables indépendamment du contexte de recherche de l'utilisateur. En effet, les requêtes exprimées par un *sac de mots clés* souffrent d'une méthode d'évaluation naïve basée sur l'appariement

requête-document. La sélection de l'information consiste à considérer que tout document contenant les mots de la requête dans n'importe quel ordre et à n'importe quel endroit de son contenu est potentiellement pertinent. Par la suite, plus le volume des ressources disponibles est important, plus la liste des résultats retournes par le système est importante. Par conséquent l'utilisateur se trouve face à une surcharge informationnelle qui le désoriente, et dans laquelle il prend la charge de distinguer ce qui est pertinent de ce qui ne l'est pas.

B. Diversité et expressivité des besoins en information

Les SRI classiques sont conçus pour servir un public le plus large possible et ne considèrent que la requête dans la représentation du besoin en information de l'utilisateur. La problématique de l'expressivité du besoin en information est d'autant plus accentué que les requêtes utilisateurs sont courtes ou ambigües.

La longueur moyenne des requêtes des utilisateurs sur le *Web* est estimée à quelques mots (environ 3 mot) Dans ce cas, les SRI renvoient à une même requête soumise par des utilisateurs différents, la même liste de résultats qui correspond à des divers centres d'intérêts ayant des estimations de pertinence comparables. Les études ont montré que l'approche généralistes des outils disponibles en RI sont à l'origine des problématiques évoqués des SRI classiques et leur incapacité à discriminer les utilisateurs en fonction de leurs centres d'intérêts ou de leurs préférences de recherche.

1.3.2 La RI adaptative

Les premières approches visant à adapter le processus de RI au contexte d'utilisation du système s'inscrivent dans le cadre de la RI adaptative. Le but fondamental de ces approches consiste à exploiter des informations additionnelles, autres que la requête, extraites des interactions de l'utilisateur avec le système, dans le but d'améliorer la recherche. Les techniques développées en RI adaptative se focalisent principalement sur l'assistance à la reformulation des requêtes des utilisateurs ou alors l'assistance à la navigation.

Nous citons les techniques de reformulation de requête par réinjection de pertinence, des techniques de désambiguïsation du sens des mots de la requête ou des techniques de regroupement thématique des résultats de recherche.

1.3.2.1 Reformulation de requêtes

Le but des techniques de reformulation de requête par réinjection de pertinence est de générer une nouvelle requête mieux adaptée au besoin en information de l'utilisateur. La reformulation de requête consiste à ajouter de nouveaux termes à une requête initiale ou alors repondérer les poids de ses termes dans le but de cibler la recherche vers les documents pertinents.

La génération de la nouvelle requête s'effectue soit par reformulation automatique ou par reformulation interactive.

La reformulation automatique de requête consiste à générer une nouvelle requête sans rétroaction explicite de l'utilisateur en exploitant les premiers documents retournés par le système en réponse à la requête initiale (*blind feedback*). La reformulation interactive des requêtes consistent à générer une nouvelle requête en exploitant un sous-ensemble de documents jugés pertinents par l'utilisateur et en réponse à la requête après chaque itération feedback.

Différents techniques de reformulation de requêtes ont été introduites dans différents modèles, notamment dans les modèles vectoriel et probabiliste.

1.3.2.2 Désambiguïsation du sens des mots de la requête

Ces techniques consistent à aider l'utilisateur d'exprimer mieux son besoin en information et orienter la recherche vers les documents portant l'intention de recherche de l'utilisateur. Elles permettent à l'utilisateur de saisir le vrai sens évoqué par les termes de la requête et l'exploiter dans des techniques d'extension des langages de requêtes. La plupart de ces techniques se basent sur l'exploitation des interfaces de clarification interactives à base d'ontologie.

L'adaptation du processus de recherche à des critères spécifiques de l'utilisateur par désambiguïsation du besoin en information derrière la requête est adopté dans plusieurs approches dans le domaine de la RI adaptative. Ces approches se basent sur la définition des paramètres mesurables à partir de la requête ou à partir du profil des premiers documents retournes par la requête. Dans ce sens, un nouveau type de réinjection de pertinence qui est le profil de requête est exploité afin de détecter des besoins en information divers et des critères de qualité de l'information derrière la requête, tels que le thème de recherche et l'information récente.

Une étude faite comporte la génération des profils temporels et thématiques des requêtes ou le profil temporel de la requête est défini par la distribution des N premiers documents retournés pour la requête en fonction de leur date de création. Cette approche permet à l'utilisateur d'affiner sa requête en choisissant d'une part le sujet d'intérêt correspondant à la requête et d'autre part de choisir ses dates préférées identifiées à partir du profil temporel de la requête afin de générer une requête ciblant les résultats estampillés selon ces dates.

1.3.2.3 Regroupement thématique des résultats de recherche

Face à la croissance du web et les difficultés rencontrées par les moteurs de recherche classiques pour satisfaire les besoins en information de l'utilisateur, les techniques de clustering/regroupement des résultats de recherche dans des catégories (Grouper (Zamir & Etzioni, 1999), Vivisimo[1], Kartoo[2], Exalead[3], etc.) ou des approches de répertorisation du web dans des taxonomies de concepts (ODP[4] ,Yahoo concept hierarchy, Google directory) ont été développées pour une accessibilité et navigation plus simple.

Les techniques de clustering sont basées sur le regroupement thématique (*clustering*) des résultats de recherche dans des catégories ou clusters à la place d'une liste de résultats paginés. Ces techniques sont basées sur le fait qu'un document qui est pertinent à une requête, a probablement une similarité avec d'autres documents qui sont peut-être aussi pertinents. Ce regroupement permet de mettre les documents similaires ensemble et avoir une idée assez générale et globale des résultats retournés et ensuite une accessibilité et une navigation plus simple. Ainsi lors d'une recherche, les utilisateurs disposeraient d'une classification virtuelle (navigation hiérarchique) des documents retournés.

Dans le même sens, plusieurs ontologies de domaines spécifiques ont été conçues et ce dans le but de faire asseoir une recherche conceptuelle permettant de simplifier la navigation à travers les catégories sémantiques de la hiérarchie utilisée.

La conception de ces ontologies consiste en des techniques de répertorisation du web permettant de classifier manuellement les pages web dans des taxonomies des concepts hiérarchiques. Il existe plusieurs répertoires du Web édités par des êtres humains, nous citons l'Open Directory Project, Yahoo concept hierarchy[5], Google directory[6] et autres.

1.3.3 Limitations de la RI adaptative

[1] Http : //clusty.com/
[2] Http : //www.kartoo.com/index.php3
[3] Http : //www.exalead.com/search/
[4] Http ://www.dmoz.org
[5] Http ://dir.yahoo.com/
[6] Http ://directory.google.com/

Malgré le gain de performance apporte par les techniques développées en RI adaptative, elles présentent toutefois des limitations. Les limites sont principalement liées à la représentation du contexte de l'utilisateur, le mode d'interaction explicite avec le système et à la dépendance de la familiarité de l'utilisateur avec le thème de recherche et le nombre d'itérations de recherche sur la performance de la RI adaptative.

- **Le contexte est peu connu :**

Les techniques de reformulation de requêtes en RI adaptative consistent à aider l'utilisateur à sélectionner les termes de la requête via des interfaces de clarification du besoin en information, ou à générer des requêtes ciblées par expansion automatique ou semi-automatique de la requête initiale. De ce fait, le contexte de recherche est limité au besoin en information véhiculé par les termes de la requête soumise à chaque itération de recherche. Le contexte dans lequel une recherche est effectuée (c'est à dire la tâche en cours, la situation géographique, etc.) est rarement utilisé pour interpréter la requête et la situer par rapport aux buts et connaissances préalables de chaque utilisateur ainsi que d'autres facteurs contextuels ayant un impact potentiel sur la performance de recherche.

- **Limitations de l'interaction explicite :**

Les techniques de reformulation de la requête par réinjection de pertinence ou de clarification du besoin en information nécessitent une rétroaction explicite de la part de l'utilisateur. Par ailleurs, plusieurs études montrent que la majorité des utilisateurs se limitent à fournir la requête initiale et préfère des mécanismes d'amélioration qui fonctionnent sans demande explicite d'information complémentaire.

- **Impact de la familiarité de l'utilisateur avec le sujet de recherche :**

La performance des techniques de reformulation de requête ou de clarification automatique du besoin en information derrière la requête est relativement dépendante de la familiarité de l'utilisateur avec le sujet de recherche d'une part et du nombre des itérations de recherche ayant un impact sur l'efficacité de la réinjection de pertinence d'autre part. Différentes études ont montré que la familiarité de l'utilisateur avec le sujet de recherche et son niveau d'expertise influent sur les performances de recherche. La sélection des documents pertinents utilisés comme sources de réinjection de pertinence dans la reformulation des requêtes diffère d'un utilisateur à un autre en fonction de sa familiarité avec le sujet de recherche. En outre,

lorsque la reformulation de requête est guidée par un contrôle de pertinence, le système devient de plus en plus opérationnel avec l'augmentation du nombre des itérations de recherche.

Ceci est accompagné par un risque de démotivation de l'utilisateur et cela d'autant plus que les qualités de rappel et précision risquent d'être faibles dans un premier temps.

Compte tenu des limitations de la RI adaptative, les approches en RI se sont orientées vers une nouvelle génération des systèmes de recherche basés sur l'accès contextuel à l'information.

1.4 La personnalisation de l'accès à l'information

La RI personnalisée guidée par le profil utilisateur est une branche de la RI contextuelle dont le contexte prend une dimension cognitive et est défini par le profil de l'utilisateur.

C'est une discipline de recherche qui est apparue avec la notion du profil utilisateur vers les années 80 avec les assistants et les agents d'interface. Le but était de créer des applications personnalisées permettant de s'adapter à l'utilisateur.

En RI personnalisée, la dimension utilisateur est décrite par son profil représentant ses centres d'intérêts, ses connaissances et ses buts de recherche.

Il a été démontré que le profil utilisateur est l'élément contextuel le plus important qui permet d'améliorer la précision de la recherche (Park., 1994). L'exploitation de cette dimension ne permet pas seulement de répondre à des requêtes ambigües ou récurrentes et les interpréter en fonction du profil utilisateur mais aussi à enrichir et faire évoluer la représentation des connaissances de l'utilisateur, ses centres d'intérêts et ses buts de recherche. Dans ce sens, la personnalisation intègre un processus de construction et d'évolution des centres d'intérêts de l'utilisateur au cours du temps.

1.5 Conclusion

Au cours de ce chapitre, nous avons présenté les principaux facteurs d'émergence de la personnalisation dans le domaine de RI, à travers l'évolution des SRI classiques vers les SRIs adaptives.

Nous avons présenté les concepts de base de la RI classique ainsi que l'évolution de la RI classique à la RI adaptative. Nous avons cité les problématiques de la RI classique en présence du contexte, les facteurs d'émergence de la RI adaptative ainsi que les techniques développées en RI adaptative, notamment la reformulation des requêtes, la désambiguïsation du sens des mots des requêtes et le regroupement thématique des résultats. Compte tenu des limitations de la RI adaptative, les travaux se sont orientées vers la RI personnalisée dont le but est de répondre mieux aux besoins en informations de l'utilisateur en exploitent les caractéristiques informationnelles spécifiques de ce dernier dans les processus d'accès à l'information.

Chapitre 2

Personnalisation de la RI et modélisation du profil utilisateur

2.1 Introduction

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

Les systèmes d'accès personnalisé à l'information visent à augmenter le processus de recherche initié explicitement par les requêtes de l'utilisateur avec des caractéristiques informationnelles comme ses centres d'intérêt, ses préférences de recherche liées à la qualité de l'information (fraicheur de l'information, genre du document, etc.) et son environnement de recherche extraites explicitement/implicitement de l'utilisateur, dans le but d'améliorer ses différents besoins.

L'ensemble de ces informations correspond à ce que l'on nomme le «contexte de l'utilisateur » ou « le profil utilisateur ».

Nous présentons dans ce chapitre la personnalisation de la recherche d'information et le profil utilisateur. Nous Commençons par aborder l'objectif général de la personnalisation et les notions de base pour la personnalisation de la recherche d'information. On entamera ensuite la modélisation de l'utilisateur pour l'accès personnalisé à l'information. Nous y présentons les principales phases du processus de modélisation : les approches de représentation, de construction et d'évolution des modèles des utilisateurs. Nous présentons, ensuite, les principaux modèles d'accès personnalisé à l'information et on passe en revue quelques systèmes développés durant ces dernières années. Enfin une conclusion générale sera présentée dans la dernière section.

2.2 L'objectif général de la personnalisation :

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

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

Dans ce contexte, (Yang & al, 2000) décrivent trois axes de réflexion.

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

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

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

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

2.3 Notions de base pour la personnalisation de la recherche d'information

2.3.1 Contexte de recherche

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

*Crestan*i et Ruthven (Crestani & Ruthven, 2007) stipulent que: «*le contexte affecte tous les aspects de la recherche d'information. Un contexte de recherche affecte la manière dont ils interagissent avec un système de récupération, quel type de réponse ils attendent d'un système et comment ils prennent des décisions concernant les objets d'information qu'ils récupèrent.* »

Dans ce cadre, le terme *contexte* se décline selon plusieurs facteurs. On ne trouve pas dans la littérature de définition complète et générique de la notion du contexte et plus précisément des éléments qui le constituent.

Les travaux de **Saracevic** (Saracevic, 1997) et **Ingerwersen** (Ingwersen, 1996) sont les premiers qui ont introduit la notion du contexte sans distinction avec la notion de situation de recherche. Le contexte y est défini selon un modèle cognitif par lequel on peut identifier des structures ou espaces cognitifs qui sont autant de variables impliquées dans le processus de RI et qui peuvent décrire les intentions et les perceptions de l'utilisateur et de ce qui l'entoure. Ces variables sont : l'espace cognitif de l'utilisateur, l'environnement social ou organisationnel, les intentions et les buts de l'utilisateur ainsi que le système lui-même.

Un contexte multidimensionnel a également été défini par **N.Fuhr** (Fuhr, 2000).Cette définition ajoute de nouvelles caractéristiques liées d'une part à l'aspect temporel du besoin en information et d'autre part au type de recherche demandé. Les trois principales dimensions retenues pour le contexte sont : social, application et temps.

- **La dimension social :** définit l'appartenance possible de l'utilisateur : individuel, groupe ou communauté.
- **La dimension application :** définit le but de la tâche accomplie : recherche ad-hoc, résolution du problème.
- **La dimension temps :** permet de définir le contexte temporel du besoin : temps passé (*batch*), intention à court terme ou intention à long terme. Le contexte à court terme (*interactif*) ou courant est associé aux besoins et préférences de l'utilisateur lors d'une session de recherche, alors que le contexte à long terme (*personnalisation*) traduit les besoins et les préférences persistants de l'utilisateur tout au long de diverses sessions de recherche.

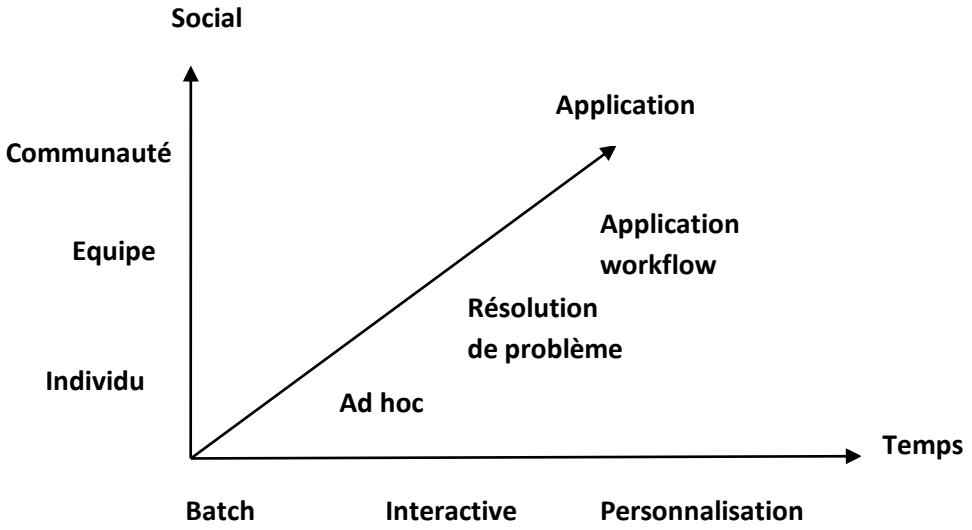

Fig.2.1 – Dimensions du contexte multidimensionnel de (Fuhr, 2000)

2.3.2Profil utilisateur

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

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

Le concept de profil utilisateur a été introduit pour l'accès à l'information en premier dans les travaux de filtrage d'information (Belkin & Croft., 1992), pour décrire une structure représentative de l'utilisateur, en l'occurrence ses centres d'intérêts. Cette notion a ensuite été ré-exploitée en RI personnalisée pour former les composantes du contexte directement dépendantes de l'utilisateur.

Il convient aussi de distinguer la notion de profil de la notion de requête. Un profil peut être défini comme une mise en équation du centre d'intérêt et des préférences de l'utilisateur, alors qu'une requête est l'expression d'un besoin circonstancié que l'utilisateur souhaite voir satisfait en tenant compte de son profil. Un profil a un caractère plus invariant que les requêtes même si le centre d'intérêt et les préférences de l'utilisateur peuvent légitimement évolué (Kobsa, 2007).

2.3.4 Architecture fonctionnelle d'un système de RI personnalisé (SRIP)

Le but fondamental d'un SRI personnalisé est de satisfaire les besoins en information de l'utilisateur en intégrant son profil dans la chaîne d'accès à l'information. L'architecture générale d'un SRIP est présentée dans la **figure 2.2**.

– L'accès personnalisé à l'information consiste alors à intégrer le profil utilisateur dans l'une des phases du processus de RI, notamment la reformulation de requêtes, l'appariement requête-document ou la présentation des résultats. Le but de cette phase est de renvoyer en haut de la liste de résultats présentés à l'utilisateur, ceux qui correspondent à ses centres d'intérêts et ses intentions de recherche.

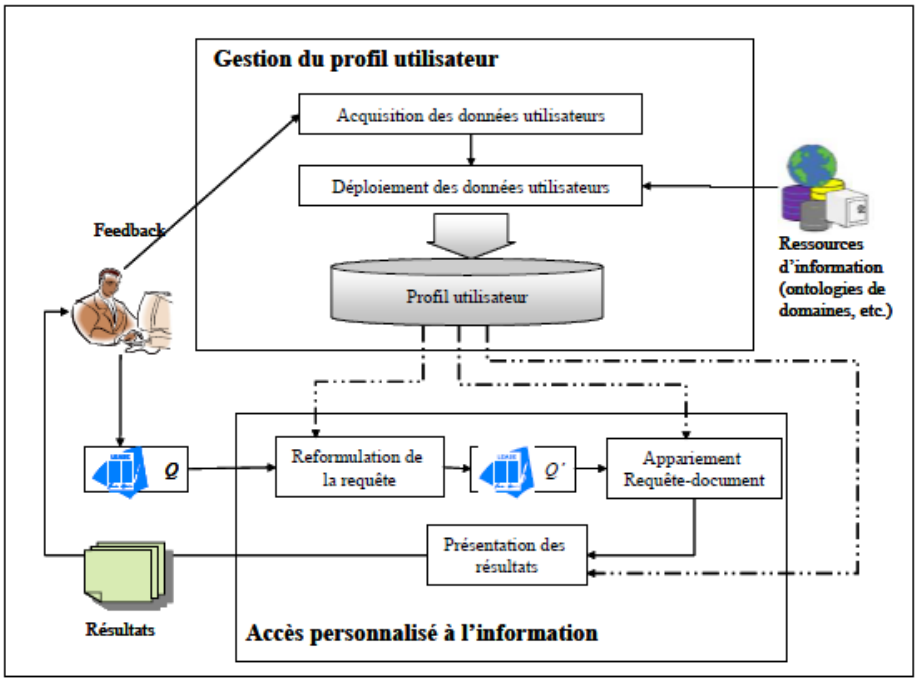

Fig. 2.2 – Architecture fonctionnelle d'un SRIP (Daoud &al,2009)

2.4 Modélisation du profil utilisateur

La modélisation de l'utilisateur est au cœur de la mise en œuvre de processus d'accès personnalisé à l'information. Elle consiste à décrire les caractéristiques informationnelles des utilisateurs à travers un modèle de profil.

Plusieurs approches et techniques ont été développées afin de modéliser l'utilisateur. Ces techniques diffèrent par leur représentation, construction et mise à jour du contenu du modèle de l'utilisateur. Ce dernier est aussi fortement dépendant du système dans lequel il évolue. En effet, généralement, les données exploitées par le système déterminent le contenu du modèle.

2.4.1 Approches de représentation du profil utilisateur

La représentation de l'utilisateur à travers la notion de profil permet de mieux comprendre certains mécanismes cognitifs, notamment ceux permettant de percevoir le concept subjectif de la pertinence et au-delà, cibler ses besoins spécifiques dans le but d'améliorer les performances de recherche. Le profil de l'utilisateur, constitué de paquets divers d'informations le caractérisant, traduit une connaissance éparse sur l'utilisateur. Dans le cadre de la RI, l'unité élémentaire utilisée pour représenter ces paquets d'informations est le **terme**

pondéré. Un modèle de représentation permet d'organiser ces éléments afin de faciliter leur exploitation dans le processus d'accès à l'information.

On distingue quatre principales approches de représentation : ensembliste, connexionniste, conceptuelle et multidimensionnelle.

- **Représentation ensembliste**

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

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

- ✓ Par une liste de mots clés, où chaque mot correspond à un centre d'intérêt spécifique. (Armstrong & al, 2005)
- ✓ Par un vecteur de termes pondérés pour chaque centre d'intérêt.(Chen & Sycara, 1998).
- ✓ Par un ensemble de vecteurs de termes pondérés (ou non) indépendants, pour prendre en compte des centres d'intérêt multiples où chaque vecteur correspond à un domaine d'intérêt. (Pazzani & al 1996)

La représentation ensembliste fut parmi les premiers modèles de profils utilisateur exploités en RI. La pondération des termes est généralement basée sur un schéma de la forme *TF*IDF* communément utilisé en RI. Le poids associé à chaque terme permet de représenter son degré d'importance dans le profil de l'utilisateur.

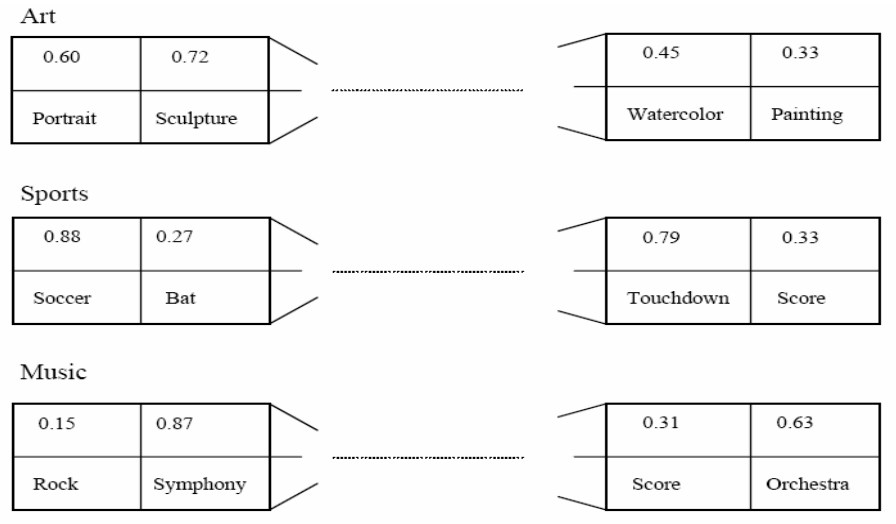

***Fig. 2.3* Un exemple de profil représenté par des mots clés. (**Zemirli & al, 2008)

La **Figure 2.3** donne un exemple de profil utilisateur représenté par des mots clés pondérés. Ce profil contient trois centres d'intérêts : *Art, Sports* et *Music*. Chaque centre est représenté par un ensemble de termes pondérés.
Music =< (*Rock*, 0.15); (*Symphony*, 0.87); … > est un extrait du l'ensemble de termes pondérés représentant le centre *Music*.

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

La représentation ensembliste du profil utilisateur apporte l'avantage de la simplicité de mise en œuvre. Cependant, même si les modèles de représentation ensembliste permettent de traduire une multiplicité des centres d'intérêts de l'utilisateur, cette représentation manque de structuration, de cohérence, des niveaux de généralité/spécificité et des relations de corrélation entre les divers centres d'intérêts de l'utilisateur.

- **Représentation connexionniste**

La représentation connexionniste du profil utilisateur consiste à représenter les centres d'intérêts de l'utilisateur par un réseau de nœuds pondérés dont chaque nœud représente un concept traduisant un centre d'intérêt de l'utilisateur. Cette représentation permet de résoudre

les failles de la représentation ensembliste par la mise en place des relations de corrélation sémantiques entre les centres d'intérêts du profil. En effet, la richesse sémantique dans cette représentation permet de résoudre le problème de la polysémie des termes inhérents à la représentation ensembliste, l'incohérence possible entre les centres d'intérêts et l'identification d'un profil adéquat au sujet de la requête via les relations sémantiques.

Plusieurs SRI personnalisés adoptent ce type de représentation. Notamment, dans ***IfWeb*** un assistant personnel à la navigation, recherche et filtrage des documents adapté aux besoins spécifiques de l'utilisateur. Le système ***Wifs,*** une interface de filtrage d'information pour personnaliser les résultats du moteur de recherche d'AltaVista. ***Infoweb,*** un système de personnalisation interactif développé pour la recherche dans les libraires digitales.

La construction d'un profil utilisateur connexionniste consiste non seulement à extraire des termes à partir des documents pertinents de l'utilisateur, mais à intégrer ces termes dans un réseau de nœuds. La construction de tels profils nécessitent la création des relations de corrélation sémantiques entre les nœuds du réseau.

La **figure2.4 (a)** illustre une description simplifiée d'un profil utilisateur hypothétique. Dans cette représentation, les termes actives, nommées planètes T_1, T_2, ..., T_n, sont contenus à la fois dans les documents jugés pertinents par l'utilisateur et dans la base de données (TDB : Tide data base créée préalablement par des experts qui sélectionnent les termes les plus pertinents pour représenter un domaine.) et les termes satellites t_1, t_2, ..., t_m sont les termes contenus dans les documents pertinents, et qui n'existent pas dans TDB, mais qui co-occurrent avec les termes T_i. La représentation abstraite d'un document est similaire à celle du profil comme illustré dans la **figure2.4 (b)**, à la seule différence que les arcs reliant les termes planètes aux termes satellites ne sont pas pondérés.

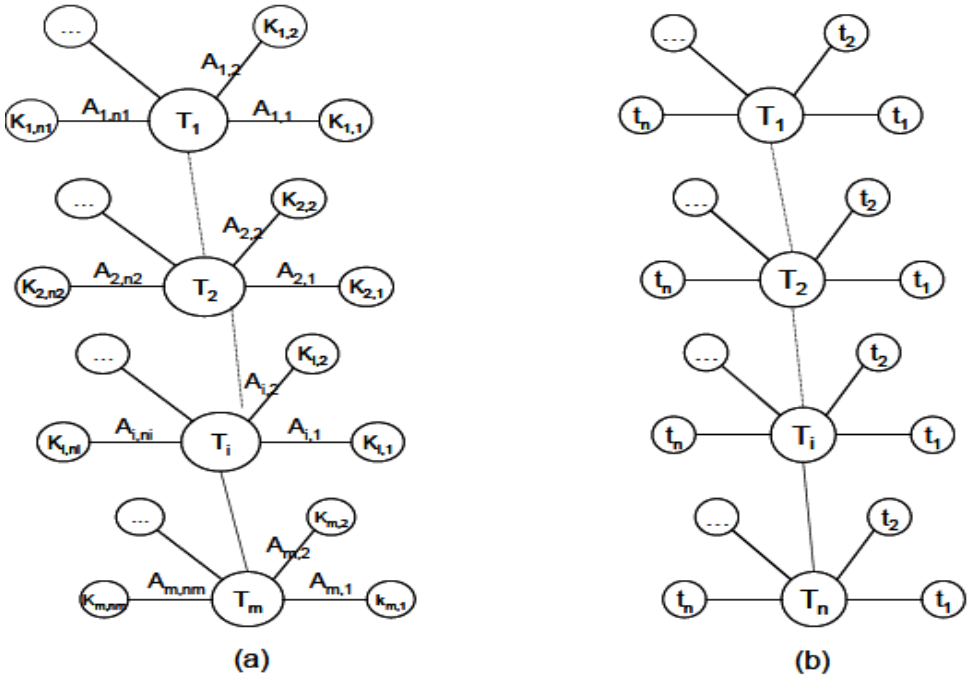

Fig.2.4 – Représentation du profil utilisateur et du document dans le système Wifs (Daoud & al, 2009)

La représentation connexionniste du profil apporte de la sémantique au modèle de l'utilisateur. Même si ce modèle de représentation permet de résoudre les problématiques liées à la représentation ensembliste en enrichissant le profil par des relations sémantiques entre les centres d'intérêts le constituant, elle présente certaines limitations. En effet, les termes du réseau sémantique représentant le profil utilisateur sont issus à partir de l'historique de recherche de l'utilisateur qui est souvent limité. Le système dispose d'un profil composé des termes avec lesquels l'utilisateur est familier. Cette situation rend la détection d'un nouveau besoin en information pour une nouvelle requête utilisateur, une tâche difficile à accomplir.

- **Représentation conceptuelle**

Dans le but de pallier les limitations de la représentation connexionniste, plusieurs études ont recours à une modélisation du profil utilisateur selon une représentation conceptuelle qui se base sur l'exploitation **des ontologies de domaines** ou **des hiérarchies de concepts**. La représentation conceptuelle consiste à représenter les centres d'intérêts de l'utilisateur par un réseau de nœuds conceptuels et reliés entre eux en respectant la topologie des liens définis dans les hiérarchies et les ontologies de domaines. Chaque concept décrivant un centre d'intérêt est représenté par un vecteur de termes pondérés où le poids traduit le degré d'intérêt de l'utilisateur dans ce concept du profil.

Il existe plusieurs hiérarchies de concepts ou des ontologies de domaines conçues dans le but de répertorier le contenu des pages web pour une navigation facile par les utilisateurs. On cite les portails en ligne tels que *Yahoo*, *Magellan*, *Lycos*, et l'*Open Directory Project (ODP)*. Ces ressources sont exploitées dans plusieurs systèmes de RI personnalisée comme une source de connaissance sémantique dans le processus de construction du profil utilisateur.

La construction d'un profil utilisateur conceptuel repose principalement sur l'utilisation de deux sources d'évidence, les données utilisateurs collectées au cours de ses sessions de recherche et des ressources sémantiques prédéfinies. L'approche de construction du profil conceptuel consiste tout d'abord à spécifier les niveaux des concepts de l'ontologie à considérer, et ensuite appliquer le procédé de déploiement des données dans des techniques de pondération de ces concepts.

Généralement, les hauts niveaux de la hiérarchie permettent de représenter le profil à long terme et les concepts de bas niveau permettent de représenter un niveau de spécificité élevé du profil utilisateur à court terme. Dans un contexte de recherche à grande échelle, telle que le Web, la représentation éparse des centres d'intérêts de l'utilisateur selon une diversité des concepts de l'ontologie présente des limitations liées à l'identification des concepts pertinents à une recherche donnée parmi une masse importante des concepts de l'ontologie représentés dans le profil utilisateur. Ceci peut augmenter considérablement le temps d'exécution des requêtes personnalisées d'une part et la gestion d'évolution du profil utilisateur d'autre part.

- **Représentation multidimensionnelle**

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

Inscrit dans une réflexion globale sur la personnalisation de l'information, une autre représentation possible du profil est la représentation multidimensionnelle. Cette

représentation a pour objectif de capturer toutes ces caractéristiques informationnelles de l'utilisateur.

Différents travaux ont abordé cet aspect sans le couvrir dans son ensemble. Ainsi, les propositions de standards P3P (W3C, 2005) 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 & Staraccia, 1999) consiste à représenter le contenu du profil utilisateur par un modèle structuré de dimensions (ou catégories) prédéfinies: *catégorie de données personnelles, catégorie de données de la source, catégorie de données de livraison, catégorie de données de comportement* et *catégorie de données de sécurité*. L'auteur a proposé ce modèle dans le cadre du développement d'un service avancé de librairie digitale (recherche et livraison personnalisées de l'information sur le *web*) : le système EUROgatherer.

Dans ce même cadre, Kostadinov (Kostadinov, 2003) a poursuivi cette classification en proposant un ensemble de dimensions ouvertes, pouvant contenir la plupart des informations susceptibles de caractériser l'utilisateur. Dans sa représentation il distingue principalement huit dimensions décrites brièvement dans ce qui suit :

- *Les données personnelles*

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

- *Le centre d'intérêt*

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*

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

✦ *La qualité attendue*

C’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*

Elle 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*

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

✦ *Le retour de préférences*

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

✦ *Les informations diverses*

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

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

2.4.2 Approches de construction du profil utilisateur

La construction du profil traduit un processus qui permet d'instancier sa représentation. L'approche de construction dépend fortement de la représentation choisie pour le profil utilisateur : les techniques utilisées par les systèmes différents selon qu'ils représentent le profil par un (des) vecteur(s) de termes ou par des classes (hiérarchiques ou pas). Cependant la démarche de construction commune à tous les systèmes est la suivante :

- on commence par collecter des informations sur l'utilisateur à partir de sources d'informations diverses,
- puis on applique des techniques et des algorithmes pour apprendre à partir de ces informations le profil de l'utilisateur.

La construction du profil s'effectue donc en deux étapes :

(1) l'acquisition et la collecte des données utilisateur ; (2) puis la construction proprement dite du profil.

(1) Acquisition des données utilisateurs

Cette phase consiste à collecter les informations pertinentes pour instancier le profil de l'utilisateur. Ce processus peut collecter ces informations soit directement à partir de la machine de l'utilisateur (côté client) ou à partir de l'application (côté serveur). Nous distinguons deux modes d'acquisition : acquisition explicite et acquisition implicite.

1. Dans l'approche explicite, les informations sont directement obtenues de l'utilisateur.
2. Dans l'approche implicite, ce sont les données de comportement de l'utilisateur qui sont exploitées.

Nous détaillons ces deux approches dans ce qui suit :

✦ Acquisition explicite

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

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

✦ Acquisition implicite

Une approche alternative remplaçant l'acquisition explicite des besoins en information de l'utilisateur, consiste à développer des algorithmes d'acquisition implicite de ces besoins. L'acquisition implicite ou « *feedback implicite* » consiste à collecter les données de l'utilisateur, en observant son comportement et en scrutant son activité. L'activité peut correspondre à :

- L'utilisation de moteur de recherche : requêtes et documents sélectionnés,
- la navigation sur le *web* : pages *web* consultées, liens sélectionnés,
- diverses applications utilisées dans le contexte de sa recherche : les applications du bureau, les outils de messagerie électronique, les éditeurs de texte, les fichiers logs,
- Consultation de bases de données ou des bases documentaires.

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

❖ Comportements observables de l'utilisateur

Un comportement observable de l'utilisateur est l'ensemble des actions qu'il effectue face aux résultats fournis par le système d'accès à l'information. L'interprétation de ce comportement sera effectuée grâce à un groupe d'indicateurs implicites (Zemirli & al, 2008).

Se basant sur les travaux de (Nichols, 1997), (Oard & Kim, 2001) proposent une catégorisation des comportements observables de l'utilisateur. Ils les classent en fonction de :

- **La catégorie de comportement** (*examiner, sélectionner, mettre en référence et annoter*), se rapporte au but fondamental du comportement observé.

- **Des unités élémentaires manipulées** (*segment, objet et classe*) se rapportent à la plus petite unité informationnelle manipulée par l'utilisateur à ce moment.

Kelly (Kelly N. J.,2004) a ajouté une cinquième catégorie de comportement, *"Création"***,** aux quatre catégorisations d'Oard et de Kim. Cette catégorie décrit les comportements que l'utilisateur lors de la création de nouvelles unités informationnelles comme par exemple l'écriture d'un papier. L'ensemble de ces classifications est regroupé, ainsi que les unités observables associées, dans le **tableau 2.1**.

L'unité sur laquelle porte l'observation

Catégorie des comportements	Segment	Objet	Classe
examiner	Regarder Ecouter Défiler Trouver Soumettre une requête	Sélectionner	Naviguer
Retenir	Imprimer	Marquer Sauvegarder Supprimer Envoyer un email	
Référencer	Copier- Coller	Répondre Ajouter un lien Citer	
Annoter	Masquer	Juger Publier	Organiser
créer	Taper, Editer	Autre	

Tableau 2.1. La catégorisation du comportement utilisateur selon (Oard & Kim, 2001),(Kelly & Teevan, 2003).

(2) Techniques de construction

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

- **Extraction d'ensemble de termes**

La technique d'extraction d'ensemble de termes est basée sur des techniques d'analyse statistique de mots clés. En effet, le contenu des documents visités par l'utilisateur est analysé pour en extraire les mots clés significatifs. Ces derniers vont servir dans l'algorithme d'apprentissage du modèle de l'utilisateur. Par exemple, dans le cadre d'une approche vectorielle, les termes extraits sont pondérés dans l'objectif de former des vecteurs de termes représentant les centres d'intérêts de l'utilisateur. Des systèmes tels que WebMate [Chen & Sycara, 98] et Alipes (Widyantoro & al., 1999), appliquent cette approche de construction.

- **Extraction de réseaux de termes**

Similairement à la technique d'extraction d'ensemble de termes, les termes sont extraits des documents jugés par l'utilisateur. Cependant, la différence réside dans la représentation des termes qui est sous forme de réseau de nœuds. Cette technique concerne principalement les représentations sémantiques du modèle de l'utilisateur. Pour construire le modèle de l'utilisateur, il est nécessaire d'exploiter des relations préexistantes entre les termes et les concepts. Ces relations peuvent se trouver dans des dictionnaires de données tels que WordNet. Des systèmes tels que SiteIF (Stefani & Strappavara., 1998) utilisent cette technique.

- **Extraction de concepts**

Cette technique de construction concerne principalement les modèles d'utilisateurs représentés par une hiérarchie de concepts pondérés. Le principe de base de cette technique est l'utilisation d'une taxonomie de concepts de référence comme profil de base. Cette dernière

peut être aussi bien l'ODP du projet Open Directory Project[1], un annuaire de concepts hiérarchique open source ou encore la hiérarchie de concepts Yahoo[2]. L'approche de construction présente de manière générale les étapes deux suivantes :

(1) identification des concepts et niveaux de l'ontologie à exploiter ;

(2) extraction des centres d'intérêts de l'utilisateur par analogie aux concepts de l'ontologie.

Des systèmes tels que **Persona** (approche de coloration d'arbre), **ARCH** (approche hybride combinant vecteurs de termes et hiérarchie de concepts) et le système du projet **OBIWAN** (association des documents collectés avec les nœuds de l'ontologie) utilisent cette technique.

2.4.3 Approches d'évolution du profil utilisateur

La gestion de l'évolution du profil utilisateur est un processus complémentaire à la construction d'un profil utilisateur et désigne leur adaptation à la variation des centres d'intérêt des utilisateurs au cours du temps. L'évolution du profil utilisateur se fait souvent selon un processus incrémental basé sur l'addition de nouvelles informations dans la représentation du profil.

Sous l'angle de la dimension temporelle, les approches gèrent l'évolution du profil à court terme, à long terme ou les deux à la fois. La gestion de l'évolution du profil utilisateur consiste principalement à capturer les changements des centres d'intérêts de l'utilisateur dans une première phase et propager ces changements au niveau de la représentation du profil.

Les techniques de collecte des informations utilisées dans la gestion de l'évolution du profil utilisateur sont relativement dépendantes de la portée temporelle du profil. On distingue le profil à court terme et le profil à long terme.

L'évolution de la représentation du profil utilisateur implique un changement des degrés d'intérêts dans certains domaines qui se traduit par une mise à jour de la structure/contenu des centres d'intérêts préalablement appris ou alors l'apparition d'un nouveau besoin en information qui se traduit par un ajout d'un nouveau centre d'intérêt au profil utilisateur.

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

- **Évolution du profil utilisateur à court terme :**

Le profil utilisateur à court terme décrit des centres d'intérêts et des besoins utilisateurs liés aux activités et la tâche de recherche courante. Souvent ces besoins en information sont partiellement représentés par le sujet de la requête.

On admet que le profil à court terme sert à mieux cibler la recherche vu qu'il contient des données considérées spécifiques et pertinentes au besoin en information courant de l'utilisateur (Shen, Tan, & Zhai, 2005). Le but fondamental de l'évolution du profil à court terme est d'améliorer la précision de recherche en utilisant le profil le plus utile et approprié à la requête et non bruité par des centres d'intérêts hors contexte de recherche. Par conséquent, ce profil permet d'adapter efficacement le processus de RI aux besoins en information spécifiques de l'utilisateur.

Dans certaines approches (Dumais & al, 2003), (Gauch & al, 2003), le profil à court terme ne représente pas nécessairement un même besoin en information mais peut traduire des multiples centres d'intérêts. Dans ces approches, l'évolution du profil utilisateur est liée à la délimitation des activités récentes de l'utilisateur par un intervalle de temps qui peut englober plusieurs sujets d'intérêts recherchés. D'autres travaux (Shen & al, 2005), (Zemirli & al, 2008), (Sieg & al, 2007) définissent le profil utilisateur à court terme dans une session de recherche par un besoin en information unique. L'évolution du profil dans ce cas requiert des mécanismes de délimitations des sessions de recherche, où une session est définie par un ensemble de requêtes liées à un même besoin en information.

- **Évolution du profil utilisateur à long terme:**

Le profil utilisateur à long terme modélise des centres d'intérêts généraux, persistants, ou récurrents de l'utilisateur et issus de son historique de recherche tout entier. Ce profil peut être exploitable dans le but d'améliorer la recherche pour toute requête soumise par l'utilisateur.

Les premiers systèmes permettant de s'adapter aux centres d'intérêts à long terme sont les systèmes de filtrage d'information tels que **Grouplens**. Plusieurs systèmes développés en RI personnalisée modélisent un profil utilisateur à long terme propre à chaque utilisateur. Parmi ces systèmes, **WebPersona** ainsi que les moteurs de recherche sur Internet **Google's Alerts**, et **Google's personalized search 1.1** et **Yahoo My Web**.

2.5 Les modèles d'accès personnalisé à l'information

La personnalisation du processus d'accès à l'information consiste à intégrer ou exploiter le profil utilisateur dans l'une des phases de processus d'accès à l'information à savoir :

(1) la phase d'appariement personnalisé de l'information,

(2) la phase de ré-ordonnancement des résultats de recherche,

(3) la phase de reformulation de requête.

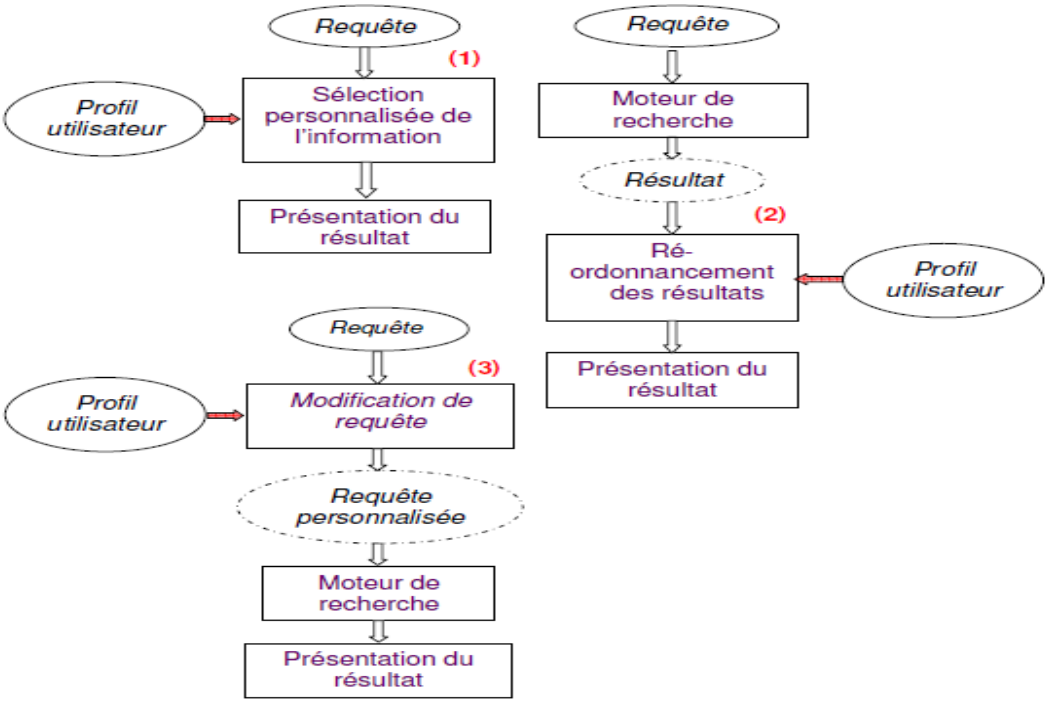

Figure 2.5. **Phases d'intégration du profil utilisateur dans le SRI** (N.Zemirli & al, 2008)

2.5.1 modèle d'appariement personnalisé de l'information

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

2.5.2 Modèle de ré-ordonnancement des résultats de recherche

La personnalisation à ce stade du processus de recherche offre une solution en réordonnant les résultats pour ne présenter à l'utilisateur que les documents pertinents en réponse à son besoin en information. Ce besoin est formulé en conjuguant les informations données par l'utilisateur tel que les requêtes soumises et celles extraites de son profil représentant ses besoins récurrents (historique, centres d'intérêts, etc.). Ainsi, la restitution des résultats s'effectue en fonction de la notion de pertinence personnelle de l'utilisateur où le rang du document est calculé en corrélation avec un utilisateur spécifique sur la base de son contexte d'interaction. Ainsi, l'idée principale du ré-ordonnancement est d'intégrer une mesure de corrélation entre le profil utilisateur et chaque document comme facteur de distinction dans le calcul du rang. (Speretta & Gauch, 2005).

$$\boldsymbol{RangFinal}(\boldsymbol{u_i}, \boldsymbol{d_j}) = \boldsymbol{\alpha} * \boldsymbol{RangConceptuel}\,(\boldsymbol{u_i}, \boldsymbol{d_j}) + (\mathbf{1} - \boldsymbol{\alpha}) * \boldsymbol{RangInitial}(\boldsymbol{u_i}, \boldsymbol{d_j}) \quad \textbf{(2.1)}$$

Tel que $\boldsymbol{RangConceptuel}(\boldsymbol{u_i}, \boldsymbol{d_j})$ est le rang du document ***dj*** obtenu en calculant une similarité entre le profil document et les concepts du profil utilisateur $\boldsymbol{u_i}$, selon la formule de similarité suivante : $\boldsymbol{sim}\,(\boldsymbol{u_i}, \boldsymbol{d_j}) = \sum_{K=1}^{N} \boldsymbol{w_t}\ (\boldsymbol{i}, \boldsymbol{k}) + \boldsymbol{W_t}(\boldsymbol{j}, \boldsymbol{k})$

Où ***N :*** le nombre totale des concepts du profil utilisateur *i* ;

$\boldsymbol{w_t(i, k)}$ **:** est le poids du concept *k* dans le profil utilisateur *i ;*

$\boldsymbol{w_t(j,k)}$ **:** est le poids du concept *k* dans le profil document *j ;*

$\boldsymbol{\alpha}$ une valeur constante entre 0 et 1 ; et $\boldsymbol{RangInitial}\ (\boldsymbol{u_i d_j})$ est le rang initialement attribué au document par le moteur de recherche.

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

2.5.3 Modèle de la reformulation de requêtes

Les requêtes utilisateur sont assurément une source évidente importante pour l'identification des besoins en information de l'utilisateur. Néanmoins, comme nous l'avons déjà mentionné, les utilisateurs soumettent souvent des requêtes très courtes et ambiguës. L'objectif de la personnalisation à ce stade du cycle de vie de la requête est de clarifier le besoin en

information de l'utilisateur en se basant sur ce que le système a appris à son sujet. Ainsi, la reformulation de requête dans ce cadre intègre les composantes informationnelles issues du profil de l'utilisateur pour identifier, enrichir et cibler son intention de recherche.

L'approche proposée par (Cui & al ; 03) se base sur le principe que, si un ensemble de documents est souvent sélectionné pour une même requête, alors les termes de ces documents sont fortement liés aux termes contenus dans la requête. Ainsi, les relations sont extraites pour déduire la distance entre les espaces de la requête et les documents. Pour cela, ils proposent d'exploiter les fichiers logs contenus dans le profil de l'utilisateur pour établir ce pont entre les deux espaces.

Ce processus débute par l'extraction d'une session de recherche pour chaque requête à partir de l'ensemble de tous les fichiers logs collectés. Chaque session est identifiée comme suit :

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

La session contient une requête ainsi que l'ensemble des documents jugés pertinents que l'utilisateur a sélectionnés (en cliquant dessus). Comme illustré par la Figure 2.6, des liens pondérés peuvent être créés entre l'espace de requête (tous les termes de la requête) et les sessions de requête et entre l'espace de document (tous les termes du document) et les sessions.

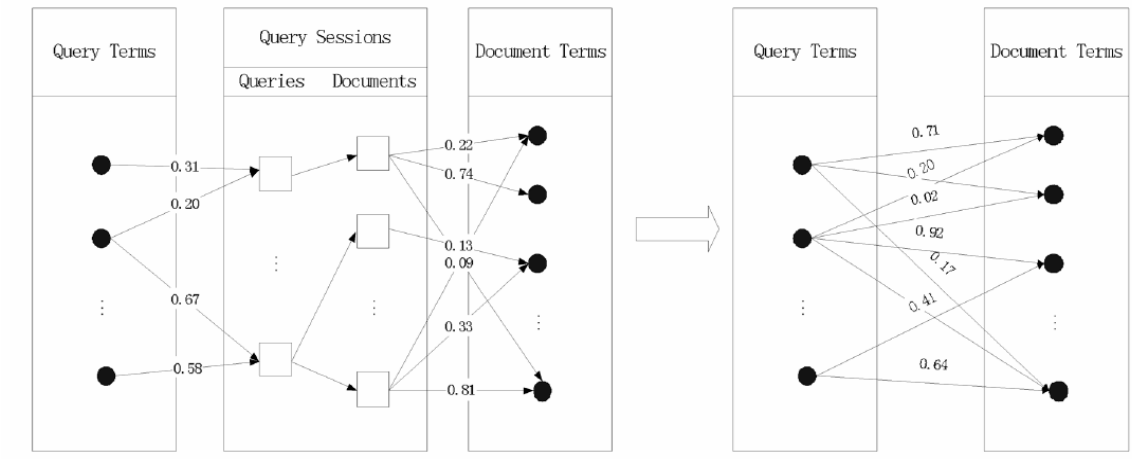

Figure 2.6. **Corrélations établies entre les termes de la requête et du document via les Sessions de requêtes (Zemirli & al , 2008)**

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

1. Extraction de tous les termes de la requête *Q*;

2. Identification des documents correspondant à chaque terme de la requête dans les sessions de requête ;

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

$$\boldsymbol{CoWeight_Q\left(w_j^{(d)}\right) = \ln\left(\prod_{w_t^{(q)} \in Q}\left(P\, w_t^{(d)}/w_t^{(q)}\right) + 1\right)} \quad (2.2)$$

Où $\boldsymbol{P\, w_t^{(d)}/w_t^{(q)}}$ est la probabilité conditionnelle mesurant le degré de corrélation entre chaque terme du document $\boldsymbol{w_j^{(d)}}$ et chaque terme requête $\boldsymbol{w_t^{(q)}}$;

4. Sélectionner *n* termes de l'espace document ayant le plus fort poids de cohésion et formuler une nouvelle requête $\hat{Q}$ en ajoutant ces termes à la requête *Q*;

5. Soumettre cette requête $\grave{Q}$ au SRI pour lancer la recherche.

2.6 Conclusion

Ce chapitre a porté essentiellement sur les principaux systèmes de personnalisation de la recherche d'information dont le point commun est la prise en compte du composant utilisateur dans le processus de recherche. Nous avons passé en revue les différentes approches et technique de modélisation du profil utilisateur, à savoir sa représentation, sa construction et son évolution au cours de temps ainsi que son intégration dans les systèmes de recherche d'information.

Nous avons abordé les principaux modèles d'accès personnalisé à l'information. Nous pouvons constater que les défis majeurs pour faire asseoir une personnalisation efficace dépendent du modèle de représentation du profil utilisateur, des mécanismes de dérivation et d'évolution du profil utilisateur au cours du temps. Ces éléments sont à la base de la différence de performance des SRI personnalisés. Compte tenu de ces éléments, l'évaluation de l'efficacité de la recherche personnalisée est aussi importante que l'évaluation de la qualité du profil appris par le système. Nous présentons dans le chapitre suivant l'évaluation des SRI et SRI personnalisés.

Chapitre 03

Evaluation des systèmes d'accès personnalisé à l'information

1. Introduction :

L'évaluation des SRI est depuis le début des travaux sur la RI un des piliers de l'évolution de ce domaine, elle consiste à mesurer les performances d'un SRI et estimer sa capacité à répondre aux besoins en information des utilisateurs. La performance ou la qualité d'un SRI est mesure en comparant les réponses du système renvoyés à l'utilisateur pour une requête donnée, aux réponses idéales que l'utilisateur espère recevoir.

L'évaluation orientée vers l'utilisateur est une composante primordiale dans le cadre de l'accès personnalisé à l'information. En effet, les objectifs d'une telle évaluation sont de mesurer l'adéquation des profils utilisateur construits par le système avec les centres d'intérêts effectifs de l'utilisateur ; ainsi que l'impact de l'intégration de ce profil, dans le processus d'accès, sur les performances de recherche.

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

2. Evaluation des systèmes d'accès à l'information

Le modèle d'évaluation Cranfield (Cleverdon, 1967) est incontestablement le modèle de référence des compagnes d'évaluation, en recherche d'information. Il est principalement fondé sur l'utilisation d'une collection de test, ou les requêtes sont les seuls ressources clés qui traduisent le besoin en information de l'utilisateur.

L'introduction de la dimension utilisateur dans le processus d'accès à l'information a cependant remis en cause la viabilité de ce modèle. Les principales limitations de ce cadre

d'évaluation sont liées à l'inadéquation des collections de test pour l'évaluation de la recherche d'information personnalisée (Kekalainen& al, 04)(L. Tamine, 2009).

Les premières tentatives faites dans ce cadre ont été proposée dans TREC à travers les tâches interactives et HARD (Herman, 03). Ces tâches ont permis l'intégration des métadonnées concernant l'utilisateur dans le processus de recherche afin d'augmenter la performance du système pour des requêtes difficiles. Les métadonnées utilisateurs concernent des critères, tels que le genre du document, la langue, etc.

Etant très spécifiques, ces tâches ne permettent pas d'évaluer un système de RI personnalisée intégrant des dimensions du contexte plus large, tel qu'un profil de utilisateur à centre d'intérêts multiples. Ceci conduit à l'émergence des approches d'évaluation fonder sur l'utilisation des contextes de recherche simulés, ou des contextes réels intégrant le profil de l'utilisateur comme étant une composante principale de la collection de test.

2.1 Le programme d'évaluation TREC

Des campagnes d'évaluation ont été mises en place au niveau mondial pour offrir un cadre standardisé et formel destiné à des protocoles d'évaluation communs. L'initiative la plus importante actuellement pour la construction de collections de tests est sans conteste TREC.
TREC est un projet international initié au début des années 90 par le NIST 3 aux Etats-Unis dans le but de proposer des moyens homogènes d'évaluation de systèmes documentaires sur des bases de documents conséquentes.
Il est co-sponsorisé par le NIST et DARPA/ITO 4. L'objectif de TREC est d'encourager les travaux de recherche d'information permettant l'accès à des bases volumineuses en fournissant:

- Une base importante de test,
- Des procédures d'évaluation uniformes,
- Un forum pour les organismes intéressés par une comparaison de leurs résultats.

2.1.1 Description d'une tâche TREC

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

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

2.1.2 Collections de test

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

1. **Les documents**

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

Les documents de cette collection proviennent de différentes sources de données : des articles de presse, des résumés courts de publications, des brevets, ainsi que des documents informatiques mis sur Internet. Il semble que certains soient (faiblement) structurés : présence

d'un titre, indication des paragraphes, les autres documents sont des structures hétérogènes annotées de métadonnées. Il existe quatre dimensions de variation :

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

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

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

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

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

Figure 3.1 : exemple d'un document de collection AP (Zemirli, 08)

La figure 3.1donne un exemple de la structure d'un document issu des articles de presse des AP 6 Newswire, collectés par AT & T Bell Laboratories pour les années 1988 et1989.

2. Topics (sujets)

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

```
<top>
<head> Tipster Topic Description
```

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

Elles sont définies par :

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

3. Les jugements de pertinence

L'évaluation est réalisée à partir d'un ensemble de documents, d'un ensemble de requêtes et d'un ensemble de jugements (liste des documents pertinents pour une requête donnée). La pertinence d'un document pour une requête est codée par une valeur numérique sur une échelle allant de non pertinent (valeur de 0) à très pertinent (valeur de +2).

Ces jugements sont regroupés dans des fichiers *Qrels*, dont la structure est la suivante :

TOPIC ITERATION DOCUMENT# RELEVANCY

Où,

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

2.1.3 Le protocole d'évaluation

Dans la plupart des protocoles, les mesures permettant l'évaluation des SRIs sont construites à partir des jugements de valeurs exprimés par des utilisateurs ou par des experts. Pour une requête et un ensemble de documents proposés en résultats, nous pouvons mesurer les taux de

performance des SRI par différentes mesures d'évaluation. Les mesures de *précision* et *rappel* sont les deux métriques les plus utilisées en RI.

- **Mesures de précision et rappel**

Les mesures de précision/rappel sont obtenues en partitionnant l'ensemble des documents, restitués par le SRI, en deux catégories : les documents pertinents et les documents non pertinents. Ces deux catégories se définissent comme suit :

-**Taux de précision** : le taux de précision mesure la capacité du système à rejeter tous les documents non pertinents à une requête. Il est donné par le rapport entre l'ensemble des documents sélectionnés pertinents *Pr* et l'ensemble des documents sélectionnés *Dr*.

$$precision = \frac{\mathrm{Pr}}{\mathrm{DR}} \qquad (3.1)$$

-**Taux de rappel** : le taux de rappel mesure la capacité du système à retrouver tous les documents pertinents répondant à une requête. Il est donné par le rapport entre les documents retrouvés pertinents *Pr* et l'ensemble des documents pertinents de la collection *R*.

$$rappel = \frac{\mathrm{Pr}}{\mathrm{R}} \qquad (3.2)$$

Idéalement on voudrait qu'un système donne de bon taux de rappel et de précision en même temps. En pratique on ne peut pas avoir un système qui aurait 100% de rappel et de précision (l'algorithme trouve la totalité des documents pertinents (rappel) et ne fait aucune erreur (précision)). Plus souvent on peut obtenir un taux de précision ou de rappel aux alentours des 30%. Notons que les deux métriques ne sont pas indépendantes. Le comportement d'un système peut varier en faveur du rappel ou de la précisons. Ainsi pour un système on a une courbe précision-rappel qui a en général la forme illustrée dans la figure 3.1

Les mesures de précision-rappel ne sont pas statiques non plus (c'est-à-dire qu'un système n'a pas qu'une mesure de précision ou de rappel). La performance du système en termes de ces mesures change en fonction de plusieurs facteurs tels que le degré de difficulté/ambigüité des requêtes de test ou autres.

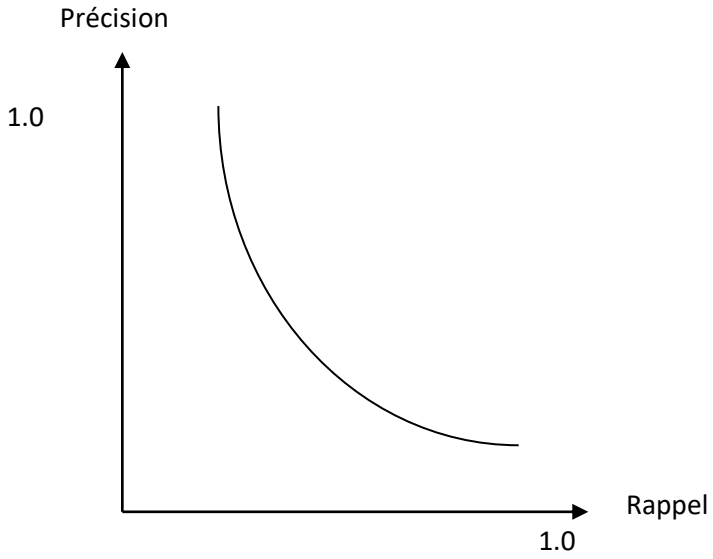

Figure 3.2 : Forme générale de la courbe de précision-rappel d'un SRI

- **Mesures à *X* documents et la précision moyenne**

Deux mesures, communément utilisées dans le cadre de TREC, sont la précision à *X* documents noté *PX* (*X* peut prendre différentes valeurs : 5; 10; 15; …, 1000) et la précision moyenne (ou *MAP* pour Median Average Precision).

La précision à *X* documents représente le nombre de documents pertinents sur les *X* premiers. Elle est souvent reliée à ce que l'on appelle la *précision exacte* ou la *R-precision*.

La précision exacte (R-precision**)** représente celle obtenue à l'endroit où elle vaut le rappel. Si la requête admet *n* documents pertinents, la précision exacte est celle calculée pour les *n* premiers documents de la liste ordonnée des documents restitués.

La précision moyenne est la moyenne des valeurs de précision à chaque document pertinent de la liste ordonnée. Elle tient compte à la fois de la précision et du rappel. Elle représente la moyenne des précisions calculées pour chaque document pertinent à trouver, au rang de ce document.

Les résultats de l'évaluation des performances du SRI sont obtenus en comparant les requêtes fournis par le système relativement à celles attendues et ce, en utilisant les mesures ci-dessus. Ainsi, chaque système restitue pour chaque requête 1000 documents, classés par ordre de

pertinence. On examine alors comment varient les indicateurs ci-dessus quand on considère les 1, 2,…, ou les 1000 premiers documents. La qualité des résultats d'un système est alors représentée par la courbe de rappel en fonction de la précision.

2.2 Limitation de l'évaluation dédié à la RI traditionnelle

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

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

2.3 Les protocoles d'évaluation pour l'accès personnalisé

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

2.3.1 Les mesures d'évaluation

Différentes mesures d'évaluation ayant été proposées dans le cadre des travaux sur la recherche des systèmes interactifs. Ces mesures peuvent être également employées pour l'évaluation d'un système d'accès personnalisé à l'information (Tamine&Calabretto, 08).

- ***la mesure RR (Relative Relevance).***

La mesure RR (Borlund&Ingwersen, 98) a pour objectif de considérer différents types de pertinence (pertinence non binaire) dans l'évaluation de l'efficacité d'un système d'accès contextuel à l'information. Cette mesure quantifie le degré de concordance entre les types de jugement de pertinence émis dans le cas de deux ensembles de jugements (soit $R1$ et $R2$)associés à une même liste de documents qui constitue les résultats d'une session de recherche. En pratique, $R1$ correspond généralement aux scores de pertinence algorithmique retournés par un SRI et $R2$ à des scores de pertinence contextuelle correspondant à un type de pertinence donné :

-situationnelle si elle est exprimée par un utilisateur,

-thématique si elle est exprimée par un assesseur etc.

La valeur de corrélation entre $R1$ et$R2$ est généralement calculée en utilisant une mesure du cosinus ; elle quantifie globalement, la capacité du système à prédire le type de pertinence contextuelle considéré.

A la différence de la mesure classique de précision, cette mesure permet de considérer les différents types de pertinence; néanmoins, elle pose un problème lors de l'évaluation comparative entre différents algorithmes de recherche voire entre différents SRI (Borlund, 03).

En effet les scores de pertinence algorithmique ne sont pas étalonnés à la même échelle entre différents SRI, ce qui rend la comparaison de mesures RR non significative.

- ***les mesure CG (Cumulative Gain) et DCG (Discount Cumulative Gain)***

Les mesures CG et DCG (Jarvelin&Kekalainen, 00), sont des mesures orientées position définies dans le contexte d'une pertinence graduelle et dont l'objectif est d'estimer le gain de l'utilisateur en termes de pertinence cumulée en observant les documents situés jusqu'à un rang donné. Ces mesures sont définies comme suit :

$$CG[i] = \begin{cases} G[1], Si \ \ i = 1 \\ CG[i-1] + G[i] \ sinon \end{cases} \qquad (3.3)$$

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

$$CG[i] = \begin{cases} G[1], Si \ \ i = 1 \\ CG[i-1] + G[i] \ /logq, sinon \end{cases} \qquad (3.4)$$

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

- ***La mesure GRP (Generalised Recall and Precision)***

La mesure GRP (Jarvelin & Kekalainen, 00) est également une mesure orientée position qui généralise les mesures classiques de rappel et précision en considérant une pertinence graduelle. Le rappel généralisé (GR) et la précision généralisée (GP) sont calculés comme suit:

$$GP = \sum_{D \in R} r(d) / \left| R \right| \qquad (3.5)$$

$$GP = \sum_{d \in R} r(d) / \sum_{d \in D} r(d) \qquad (3.6)$$

Où *R* est l'ensemble des documents retournés par le SRI,
D est l'ensemble des documents de la collection,
r(*d*) est la valeur de pertinence graduelle associée au document *d*.

De manière analogue aux mesures classiques de rappel/précision, ces mesures offrent la possibilité d'être agrégées pour plusieurs requêtes ou plusieurs niveaux de rappel et donnent ainsi la possibilité de tracer des courbes de performances.

2.3.2 Collection de test

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

1. réutilisation des collections de test de TREC (documents, requêtes et jugements de pertinence) puis leur augmentation par des éléments du contexte. Ces éléments, tel que l'historique des interactions, sont extraits à partir des interactions d'utilisateurs effectifs interrogeant la base TREC à l'aide de requêtes TREC (Shen& Tan &Zhai, 05). Le référentiel d'évaluation étant disponible, les mesures agrégées de rappel/précision sont alors exploitées pour évaluer les différences de performances entre le scénario de recherche basique (ne tenant pas compte du contexte) et le scénario de recherche contextuelle.

2. construction de collections de test en menant une campagne d'évaluation avec des utilisateurs réels : c'est le protocole adopté par la plupart des travaux. (Teevan, 05).

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

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

chaque utilisateur spécifique et pour chaque requête. L'ensemble des documents jugés constitue la collection de référence.

2.3.3 Scénarios d'évaluation d'un SRIP

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

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

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

Etape 1 : Evaluer la qualité des profils appris. Lors de cette étape, la qualité du profil se traduit par son adéquation avec les centres d'intérêts effectifs de l'utilisateur. Pour cela, un découpage de la collection de test est effectué en deux sous-collections : une sous-collection pour l'apprentissage du profil utilisateur et une sous-collection pour les tests à effectuer. Et ensuite, ces tests peuvent être effectués en utilisant des mesures quantitatives.
Ces mesures permettent de quantifier le degré de précision des profils construits relativement aux annotations explicites des utilisateurs (Chaffee&Gauch, 00) ; (Dumais & al, 03). En outre, cette étape peut inclure des tests pour évaluer l'efficacité de l'algorithme d'apprentissage du profil. Dans ce cas, des mesures comparatives entre plusieurs algorithmes (Pazzani& al, 96) peuvent être utilisées où des mesures de convergence de l'algorithme (Liu &Yu, 04).

. **Etape 2** : Validation de l'accès personnalisé.

L'objectif de cette étape est de tester l'amélioration des performances de la recherche. Les scénarios expérimentaux consistent, de manière classique, à comparer les performances de recherche d'un moteur de recherche classique (sans intégration du profil) et du moteur de

recherche personnalisé proposé intégrant le profil de l'utilisateur (Liu &Yu, 04) ; (Speretta&Gauch, 04) ; (Gauch& al, 03).

3. Conclusion :

Ce chapitre est consacré aux approches d'évaluation des systèmes d'accès personnalisé à l'information. Des différents éléments abordés dans cette section, il en ressort deux points importants : les campagnes d'évaluation standard largement utilisées en RI tel que TREC, ne sont pas adaptées à la RI personnalisée. Ces protocoles d'évaluation sont centrés requête et non utilisateur.

L'évaluation orientée « utilisateur » est une composante primordiale dans le cadre de l'accès personnalisé à l'information. Les objectifs d'une telle évaluation sont de mesurer l'adéquation des profils utilisateur construits par le système avec les centres d'intérêts effectifs de l'utilisateur ; ainsi que l'impact de l'intégration de ce profil, dans le processus d'accès, sur les performances de recherche. Néanmoins, une telle évaluation reste une problématique majeure dans le domaine de la recherche d'information personnalisée.

Dans le chapitre suivant nous allons présenter notre démarche de création d'une collection de test intégrant la dimension profil utilisateur afin d'évaluer son impact sur le processus de recherche.

Chapitre 04

Evaluation pour un accès personnalisé à l'information.

4.1 Introduction

La modélisation de l'utilisateur est au cœur de la mise en œuvre de processus d'accès personnalisé à l'information. Elle consiste à décrire les caractéristiques informationnelles des utilisateurs à travers un modèle de profil et cela afin de lui offrir une meilleure réponse à ses besoins en information, en tenant compte de ses centres d'intérêts.

Le chapitre 2 présenté dans ce mémoire, a permis de cerner le domaine de la recherche d'information personnalisée, ainsi que la notion du profil utilisateur autour de laquelle s'articule ce dernier.

Notre contribution présentée dans ce chapitre, porte sur l'évaluation de l'impact d'intégration du profil utilisateurs dans la recherche d'information par l'utilisation des profils des utilisateurs réels.

Le profil de l'utilisateur est construit d'une manière implicite. Les sources d'information utilisées pour sa définition sont extraites à partir des données issues des interactions de l'utilisateur avec le système de recherche. Il est défini par ses centres d'intérêts, organisés selon des structures ensemblistes, basées sur des vecteurs de termes pondérés.

Pour la personnalisation du processus de recherche, notre approche consiste à intégrer le profil de l'utilisateur dans la phase d'appariement pour calculer la pertinence des documents, en fonction des caractéristiques spécifiques de l'utilisateur. L'idée est de substituer à la fonction de pertinence classique qui mesure le degré d'appariement requête-document, une fonction requête-document-profil de l'utilisateur (Achemoukh , 2018).

Nous abordons au cours de ce chapitre la problématique et les motivations qui ont suscité la mise en place de notre approche d'évaluation.

4.2 Démarche d'évaluation

Les approches proposées en recherche d'information personnalisée sont confrontées à la question de la définition des informations nécessaires concernant l'utilisateur et la gestion d'évolution de son profil au cours du temps ainsi que son exploitation dans le processus de recherche.

Notre objectif dans le cadre du domaine de la personnalisation consiste à la mise en place d'une collection de test personnalisée intégrant des profils des utilisateurs réels afin d'évaluer leur impact sur la RI.

Le scénario d'évaluation que nous avons suivi est effectué avec 7 utilisateurs (étudiants de l'université de Mouloud MAMMERI de Tizi-Ouzou).Chaque utilisateur soumet entre 3 à 4 requêtes à Google. Durant ces recherches on collecte pour chaque paire (utilisateur, requête) les 10 premiers résultats retournés par le moteur de recherche, ainsi que les différentes informations liées aux interactions utilisateur (la sélection, le temps de lecture, le référencement, la sauvegarde des documents).

Nous avons utilisé le langage XML pour structurer les informations brutes de notre collection, qui est un langage informatique de **balisage générique**, il permet de structurer de manière hiérarchisée et organisée les données d'un document. La structure générale du fichier XML pour chaque utilisateur est illustrée dans la figure 4.2.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<user >
< !- -  la première interaction de l'utilisateur - - >
<interaction >
< !- -  La requête soumise par l'utilisateur lors de son interaction avec le système de recherche  - - >
<query></query>
< !- -  les 10 premiers résultats retournés par le système en réponse à la requête utilisateur - - >
<results>
< !- -  le 1er résultat retourné par le système en réponse à la requête et les factures de jugement de pertinence - - >
<result>
<url></url>
<title></title>
<snippet> </snippet>
</result>
< !- -  le 10eme  résultat retourné par le système en réponse à la requête et les factures de jugement de pertinence - - >
<result>
<url>L'adresse de document</url>
<title>le titre de document</title>
<snippet>une  description de contenu de document </snippet>
</result>
</results>
</interaction>
</user>
```

La figure 4.1 : la structure générale des interactions utilisateur avec XML

La figure 4.2 illustre un exemple d'un extrait des interactions de recherche de l'utilisateur N°2 :

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<user num="1"  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"xsi:noNamespaceSchemaLocation="user_1.xsd"  >
<interaction num="1">
<query>mémoire sur la gestion des déchets</query>
<results>
<result  rank="1" TL="1" examiner="1" retenir="1"  référencer="0" >
<url>http://www.univbejaia.dz/dspace/bitstream/handle/123456789/%A9jaA.pdf?sequence=1&amp;isAllowed=y</url>
<title>Contribution à l'étude des déchets ménagers de la ville de Béjaia par cartographie numérique</title>
<snippet>La présente étude porte sur l'utilisation des Systèmes d'information Géographiques
        (SIG) comme outil d'aide à la gestion des déchets ménagers de la ville de Bejaia.</snippet>
</result>
      .....
<result rank="10"  TL="0"  examiner="0"  retenir="0"    référencer="0">
<url>http://bibfac.univ-tlemcen.dz/snvstu/opac_css/doc_num.php?explnum_id=2611</url>
<title>Identification et Caractérisation des déchets ménagers solides de la ville de Tlemcen</title>
   <snippet>Le problème des déchets solides en Tlemcen se pose avec acuité, La production de déchets est liée à la population,
                                                    L'augmentation...</snippet>
</result>
</results>
</interaction>
…
</user>
```

Figure 4.2 : un extrait des interactions de recherche de l'utilisateur N°2.

Afin de récupérer les données du fichier XML et calculer une mesure de pertinence pour chaque document, nous avons utilisé l'implémentation Java de l'API DOM (Document Object Model) qui est une API inter-langage du World Wide Web Consortium (*W3C*) pour accéder et modifier les documents XML.

La mesure de pertinence calculée permet de sélectionner les Tops documents pour chaque requête qui vont former notre collection, qui sera utilisé dans la construction des profils utilisateurs.

Cette collection est découpée en 2 sous-ensembles :

- Une collection d'apprentissage qui regroupe les Top documents de 2 requêtes choisi au hasard afin de définir les profils correspondants.
- Une collection de test qui regroupe les Top documents d'une 3éme requête restante afin de personnaliser sa recherche.

Pour l'évaluation de l'impact d'intégration du profil dans le processus d'accès, nous avons effectué une comparaison entre les résultats de la recherche classique et les résultats obtenus après l'intégration ce profil lors de l'évaluation de la 3éme requête.

4.2.1 Représentation et construction du profil

Dans notre cas, la construction du profil utilisateur consiste à utiliser les tops document Dr d'une requête d'apprentissage, le profil u est alors représenté sous forme d'un vecteur de termes pondérés mesurés comme suit :

$$W_{ti,u} = \frac{\sum_{d_j \in Dr} \left(\frac{tf_{t_i,d_j}}{\sum_{t_i} tf_{t_i,d_j}} \right)}{|Dr|} \qquad \textbf{(4.1)}$$

$W_{t,u}$: Le poids du terme t_i dans le profil u.

$freq_{t_i,d_j}$: La fréquence du terme t_i dans le document d_j.

$\sum freq_{t_i,d_j}$: La somme des fréquences des termes t_i dans le document d_j.

$|Dr|$: Le nombre des tops documents.

4.2.2 Illustration de notre approche

Pour illustrer notre approche nous avons pris l'exemple du premier utilisateur (user1), qui a effectué un ensemble d'interactions.

- **Interaction 1 :**

Cette interaction est caractérisé par la requête Q1= {régulateur de pression}.

Le système a retourné la liste des document suivante : D={d1,d2,d3,d4,d5,d6 d7,d8,d9,d10}.(respectivement dans l'ordre de classement de Google)

Soit Dr= {d1} le document jugés pertinent par l'utilisateur lors de cette interaction.

L'index du document d1 inclut 10 termes, dont le nombre d'occurrences de chaque terme pour la requête Q1 et le document d1 est exprimé dans le tableau 4.1 suivant :

	Alimentation	Eau	Régulateur	Réducteur	Pression	Produit	Compteur	Manomètre	Testeur	Filtre
Q1	0	0	1	0	1	0	0	0	0	0
d1	5	14	14	7	23	11	5	8	2	3

Tableau 4.1 : Fréquence des termes dans la requête et dans le document d 1.

Le tableau 4.2 présente un récapitulatif des poids des termes du document d1, obtenue selon la formule (4.1).Ce vecteur représente le centre d'intérêt (profil) inféré au cours de l'interaction1.

	Alimentation	Eau	Régulateur	Réducteur	Pression	Produit	Compteur	Manomètre	Testeur	Filtre
VQ1	0 ,054	0 ,152	0 ,152	0,076	0 ,250	0,119	0,054	0,086	0,021	0,032

Tableau 4.2 : Poids des termes dans les Tops documents de la requête 1

- ✓ Le document d1 contient des termes rares comme le terme « testeur » et « filtre », et un grand nombre de terme qui apparaissent souvent comme « pression », « régulateur » et « eau » donc les mieux pondéré et le plus représentatifs du contenu du document d1 .

- **Interaction 2 :**

  Cette interaction est caractérisé par la requête Q2= {régulateur de pression pneumatique}. Le système a retourné la liste des documents suivante : D= {d1, d2, d3, d4, d5, d6 d7, d8, d9, d10} (respectivement dans l'ordre de classement de Google)

Soit Dr= {d6, d1} le document jugés pertinent par l'utilisateur lors de cette interaction.

L'index des tops document inclut 10 termes, dont le nombre d'occurrences de chaque terme pour la requête Q2 et pour chaque document est exprimé dans le tableau 4.3 suivant :

	Pression	Régulateur	Air	Raccord	Nelson	Système	Pneumatique	Flou	Valve	Universel
Q2	1	1	0	0	0	0	1	0	0	0
d6	18	17	0	15	7	5	0	9	0	6
d1	12	7	23	0	0	6	9	0	8	0

Tableau 4.3 : Fréquence des termes dans la requête et les documents.

Le tableau 4.4 présente un récapitulatif des poids des termes des tops documents de la requête Q2, obtenue selon la formule (4.1).Ce vecteur représente le centre d'intérêt (profil) inféré au cours de l'interaction 2.

	Pression	Régulateur	Air	Raccord	Nelson	Système	Pneumatique	Flou	Valve	Universel
VQ2	0 ,209	0 ,164	0 ,176	0,097	0 ,057	0,078	0,069	0,058	0,061	0,038

Tableau 4.4 : Poids des termes dans les Top documents de la requête 2

- ✓ Les documents d1 et d6 sont proche en contenu ils contient un nombre de terme partager comme « Pression» et « Régulateur » et des termes différent qui apparaissent souvent dans chacun des document comme « raccord » pour d1 et « Air » et « Pneumatique » pour d6. Ces termes sont les mieux pondéré et les plus représentatifs du contenu des deux documents.

➢ **Interaction 3 :**

Cette interaction est caractérisé par la requête Q3= {régulateur de pression gaz naturel}.Le système a retourné la liste des documents suivante : D= {d1, d2, d3, d4, d5, d6, d7, d8, d9, d10} (respectivement dans l'ordre de classement de Google)

Soit Dr= {d4, d2} le document jugés pertinent par l'utilisateur lors de cette interaction.

L'index des tops document inclut 10 termes, dont le nombre d'occurrences de chaque terme pour la requête Q3 et pour chaque document est exprimé dans le tableau 4.5 suivant :

	Pression	Régulateur	Gaz	Série	Etagé	Naturel	Soupape	Surpression	Protection	Décompression
Q3	1	1	1	0	0	1	0	0	0	0
D4	54	72	9	0	0	0	10	9	9	8
D2	71	46	41	30	27	17	0	0	0	0

Tableau 4.5 : Fréquence des termes dans la requête et les documents.

Le tableau 4.6 présente un récapitulatif des poids des termes des tops documents de la requête Q3 obtenue selon la formule (4.1).Ce vecteur représente le centre d'intérêt (profil) inféré au cours de l'interaction 3 :

	Pression	Régulateur	Gaz	Série	Etagé	Naturel	Soupape	Surpression	Protection	Décompression
VQ3	0 ,299	0 ,338	0 ,108	0,060	0 ,054	0,034	0,029	0,026	0,026	0,023

Tableau 4.6 : Poids des termes dans les Top documents de la requête 3

✓ Les documents d2 et d4 sont proche en contenu ils contient un nombre de terme partager comme « Pression» et « Régulateur » et « gaz » et des termes différent qui apparaissent souvent dans chacun des document comme « série », « Etagé » et « naturel » pour d2 et « Soupape »et «surpression » pour d6. Ces termes sont les mieux pondéré et les plus représentatifs du contenu des deux documents.

- Ensuite On considère la requête Q2 et Q3 des requêtes d'apprentissage des centres d'intérêts dans le but de construire le profil utilisateur, la requête Q1 est utilisé pour le test.

Requêtes d'apprentissage	Requête de test
Q2, Q3	Q1

Le profil appris à partir des requêtes d'apprentissage est utilisé dans la personnalisation du processus de recherche de la requête de test.

1. Construction de profil de l'utilisateur 1 :

Le profil est construit à partir de la moyenne des poids des termes partager entre les deux sous profils de Q2 et Q3, et des termes les mieux pondéré. Le vecteur profil résultat est définis comme suit:

	Pression	Régulateur	Air	Gaz	Raccord	Valve	Série	Système	Pneumatique	Flou
VP_G	0,359	0 ,333	0 ,176	0,108	0 ,097	0,061	0,060	0,078	0,069	0,058

Tableau 4.7 : Poids des termes du profil utilisateur 1

2. la personnalisation du processus de recherche :

Le profil appris est alors intégrer dans le processus de recherche, puis un test est effectuer avec la requête Q1= {régulateur de pression} pour évaluer l'impact de cette intégration sur les résultats de recherche.

- Lors de la recherche classique l'utilisateur a soumet la requête Q1= {régulateur de pression}, les résultats retournée portent sur « les régulateurs de pression hydraulique ». Selon VQ1={ Alimentation 0 ,054, Eau 0 ,152, Régulateur 0 ,152, Réducteur 0,076, Pression 0 ,250, Manomètre 0,119 ; Produit 0,054, Compteur 0,086, Testeur 0,021, Filtre 0,032}.

  Cependant, selon le profil construit VP_G= {Pression 0,359 ; Régulateur 0,333 ; Air 0,176 ; Gaz 0,108 ; Raccord 0,097 ; Valve 0,061 ; Série 0,060 ; Système 0,078 ; pneumatique 0,069 ;Flou 0,058 } le besoin en information de l'utilisateur porte sur « les régulateur de

pression pneumatique » alors la personnalisation de processus de recherche en intégrant ce profil va retourner à l'utilisateur en réponse à la requête Q1 des résultat portant sur « les régulateur de pression pneumatique » et non pas sur « les régulateur de pression hydraulique ».

4.3 Conclusion

Dans ce chapitre, nous avons présenté notre contribution portant sur l'impact de l'intégration du profil utilisateur dans la recherche d'information par l'utilisation des profils des utilisateurs réels.

Pour cela nous avons commencé par construire une collection de test réel utilisé dans la construction des profils qui sont ensuite intégré dans le processus de recherche afin de personnaliser l'accès à l'information.

Notre résultat obtenu lors de la recherche personnalisée comparativement au résultat de la recherche classique révèle une performance de recherche plus élevé, vérifiant ainsi notre approche de construction et d'exploitation de profil utilisateur dans le processus de recherche.

CONCLUSION GENERALE

Conclusion générale

Le travail présenté dans ce mémoire s'inscrit dans le cadre de la personnalisation de l'information dans le domaine de la RI. Nous avons alors créé une collection de test réel, à partir de laquelle nous avons construit des profils et enfin nous avons vérifié l'impact de l'intégration de ces profils dans le processus de recherche.
Ce travail peut se résumer en deux parties :

- Dans la première partie (chapitre 1 et chapitre 2 chapitre 3), nous avons fait une étude théorique sur les concepts de base de la recherche d'informations classiques, puis la présentation de la recherche d'informations personnalisées ainsi que la problématique liée à la mise en place d'une compagne d'évaluation et formelle pour l'accès personnalisé.

- Dans la deuxième partie (chapitre 4) nous avons créé une collection de test personnalisé que nous avons utilisé pour la définition et la construction des profils utilisateurs puis on a discuté les résultats de l'évaluation de l'impact d'intégration de ces profils dans la recherche d'informations.

Le présent travail nous a permis de voir l'importance de la recherche personnalisée et de comparer entre la recherche classique et recherche personnalisée en RI qui permet réellement de sélectionnée l'information pertinente répondant aux besoins d'un utilisateur.

Comme perspectives à ce travail, nous envisageons :

- Enrichir la collection avec de nouveaux utilisateurs et plus de documents collection avec plus d'utilisation.
- implémentation d'un processus de recherche intégrant les profils des utilisateurs apprit à partir de la collection réelle.

Références Bibliographique

Achemoukh F, Ahmed-Ouamer R. (2018), "Prise en compte du profil utilisateur en recherche d'information selon l'approche bayésienne". ISKO- Maghreb 3rd International Symposium, pp: 1-7.

Amato, G., & Staraccia, U. (1999). User profile modelling and applications to digital librairies. *In Proceedings of the 3rd European Conference on Research and avanced technology for digital libraries*, (pp. 184-187).

Armstrong, R., Freitag, D., Joachims, D., & Mitchell, T. (2005). Webwatcher : A learning apprentice for the world wide web. *In Spring symposium on Information gathering from Heterogeneous, distributed environments*, (pp. 6-12).

Belkin, N., & Croft, W. (1992). Two sides of the same coin ? *Communication of the ACM, 35(12)*, pp. 29-38.

H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma. Query expansion by mining user logs. *IEEE Trans. Knowl. Data Eng.*, 15(4) :829–839, 2003.

Chen, L., & Sycara, K. (1998). Webmate : A personal agent for browsing and searching. *In Proceedings of the 2nd international conference on autonomous agents and multiagent systems*, (pp. 10-13). Minneapolis.

C. Cleverdon. The cran_eld test on index language devices. *Aslib*, 19(6) :173.194,1967.

Crestani, F., & Ruthven, I. (2007). Introduction to special issue on contextual information retrieval systems. Information Retrieval., *10(2)* , pp. 111-113.

Daoud, M., Boughanem, M. M., & Tamine-Lechani, L. (2009). *Accès personnalisé à l'information approche basée sur l'utilisation d'un profil utilisateur sémantique dérivé d'une ontologie de domaines à travers l'historique des sessions de recherche.* Thése de doctorat , Université Paul Sabatier de Toulouse, Toulouse.

Dumais, S., Cadiz, E. C., Jancke, G., Sarin, R., & Daniel, C. R. (2003). Stuff i've seen : a system for personal information retrieval and re-use. *In SIGIR'03 : Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 72–79). ACM Press.

Dumais, S., Cadiz, E. C., Jancke, G., Sarin, R., & Daniel, C. R. (2003). Stuff i've seen : a system for personal information retrieval and re-use. *In SIGIR'03 : Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 72–79). ACM Press.

Fuhr, N. (2000). Information retrieval : introduction and survey. *post-graduate course on information retrieval, university of Duisburg-Essen.* Germany.

Gauch, S., Chaffee, J., & Pretschner, A. (2003). Ontology-based personalized search and browsing. *Web Intelligence and Agent Systems, 1(3-4)* , pp. 219-234.

Ingwersen, P. (1996). Cognitive perspectives of information interactions : Elements of a cognitive ir theory. *Annual review of information science and technology, 52(1)* , pp. 3-50.

J. Kekalainen and K. Jarvelin. Evaluating information retrieval systems under the challenges of interaction and multidimensional dynamic relevance. In Proceedings of the 4th CoLIS conference, pages 253–270. P. Ingwersen and P. Vakkari, 2004.

Kelly, D., & Teevan, J. (2003). Implicit feedback for inferring user preference : a bibliography. *SIGIR Forum, 37(2)* , pp. 18-28.

Kelly, N. J. (January 2004). Understanding implicit feedback and document preference : a naturalistic study. *In PHD dissertation. Ritgers University* . New Jersey.

Kobsa, A. (2007). Privacy-enhanced web personalization. In The Adaptive Web : Methods and Strategies of Web Personalization, Lecture Notes in Computer Science. Dans K. P. In Brusilovsky (Éd.). *4321*. Berlin Heidelberg New York: Springer-Verlag.

K. Jarvelin and J. Kekalainen. Ir evaluation methods for highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 41.48. Belkin and al Eds, 2000.

Liu, F., Yu, C., & Meng, W. (2004). Personalized web search for improving retrieval effectiveness. *IEEE Transactions on Knowledge and Data Engineering, 16(1)* , pp. 28–40.

Nichols, D. (November 1997). Implicit rating and filtering. . *In In Proc. 5th DELOSWorkshop on Filtering and Collaborative Filtering*, (pp. 31-36). Budapest, Hungary.

Oard, D. W., & Kim, J. (2001). Modeling information content using observable behavior. *In Proceedings of the 64th Annual Meeting of the American Society for InformationScience and Technology*, (pp. 38-45). USA.

Pazzani, M. J., Muramatsu, J., & Billsus, D. (1996). Syskill & webert : Identifying interesting web sites. *In Proceedings of the 30th National Conference on Artificial Intelligence*, (pp. 54-61). Portland.

Park., T. K. (1994). Toward a theory of user-based relevance : a call for a new paradigm of inquiry. *Journal of the American Society for Information Science , 45* (3), 135–141.

Research, I. A., Zien, J., Meyer, J., & Tomlin., J. (2001). Web query characteristics and their implications on search engines jason zien, j org meyer, john tomlin. *In Proceedings of the 10th International WWW Conference.* Hong Kong.

Robertson, S., & Jones, K. S. (1976). Relevance weighting for search terms. *Journal of The American Society for Information Science , 27* (3), 129-146.

S. Speretta and S. Gauch. Personalizing search based user search histories. In *Proceedings of the 13th International Conference on Information Knowledge and Management*, pages 238.239, 2004

Salton, G. (1971). The SMART Retrieval System-Experiments in Automatic Document Processing. *Prentice- Hall Inc* .

Salton, G., & Yang, C. (1973). *Journal of documentation* , 351-372.

Salton, G., & McGill., M. (1983). Introduction to Modern Information Retrieval. McGraw-Hill, New York.

Saracevic, T. (1997). The stratefied model of information retrieval interaction : extension and applications. *In Proceedings of the 60th annual meeting of the American Society for Information Science*, (pp. 313-327). Medford.

Shen, X., Tan, B., & Zhai, C. (2005). Implicit user modeling for personalized search . *In CIKM '05 : Proceedings of the 14th ACM international conference on Information and knowledge management* (pp. 824–831). New York, NY, USA: ACM.

Sieg, A., Mobasher, B., & Burke, R. (2007). Web search personalization with ontological user profiles. *In CIKM'07 : Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (pp. 525–534). New York, NY, USA: ACM.

Speretta, M., & Gauch, S. (2005). Personalized search based on user search histories. *In WI '05 : Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 622–628). Washington, DC,USA: IEEE Computer Society.

Stefani, A., & Strappavara., C. (June 20-24 1998). Personalizing access to web sites : The siteif project. *In Proceedings of the 2nd Workshop on Adaptive Hypertext and Hypermedia.* Pittsburgh.

Tamine, L., Boughanem, M., & Daoud, M. (2009). Evaluation of contextual information retrieval effectiveness : overview of issues and research. *Knowledge and Information Systems.*

Tamine-Lechani, L., Boughanem, M., & Zemirli, N. (2008). Personalized document ranking : exploiting evidence from multiple user interests for profiling. *In Journal of Digital Information Management* .

Teevan, J., Dumais, S., & Horvitz, E. (August 15-19 2005). Personalizing search via automated analysis of interests and activities. *In Proceedings of the 28th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 449-456). Salvado,rBrazil.

W3C. (2005).

Widyantoro, D., Yin, J., Nasr, M. E., Yang, L., Zacchi, A., & Yen., J. (March 22-24 1999). Alipes : A swift messenger in cyberspace. *In Proceedings of Spring Symposium Workshop on Intelligent Agents in Cyberspace.* Stanford.

Yang, Q., H. F. Wang, G., Zhang, J. R., Lu, Y., Lee, K. F., & Zhang, H. J. (2000). Toward a next-generation search engine. *In Proceedings of the Sixth Pacific International Conference on Artificial Intelligence* (pp. 5-15). Melborne, Australia: Springer.

Zamir, O., & Etzioni, O. (1999). Grouper : a dynamic clustering interface to Web search results. *Computer Networks : The International Journal of Computer and Telecommunications Networking, 31(11)*, pp. 1361–1374.

Zemirli, N., BOUGHANEM, M. M., & L.Tamine-Lechani. (2008). *Modèle d'accès personnalisé à l'information basé sur les Diagrammes d'Influence intégrant un profil utilisateur évolutif.* Toulouse: Université Paul Sabatier de Toulouse III.

Zemirli, W. N., Tamine, L., & Boughanem, M. (2005). Accès personnalisé à l'information :vers la définition d'un profil utilisateur multidimensionnel. *In International Symposium On Programming Systems* (pp. 20-28). USTHB.