République **A**lgérienne **D**émocratique et **P**opulaire
Ministère de l'**E**nseignement **S**upérieur et de la **R**echerche **S**cientifique

UNIVERSITE **M**OULOUD **M**AMMERI DE **T**IZI-**O**UZOU

FACULTE DU **G**ENIE **E**LECTRIQUE ET D'**I**NFORMATIQUE
DEPARTEMENT D'INFORMATIQUE

Mémoire de Fin D'Etudes de MASTER ACADEMIQUE

Domaine : **Mathématiques et Informatique**
Filière : **Informatique**
Spécialité : **Systèmes informatiques**

Présenté par :

MESSARA SABRINA
DOUAL KARIM

Thème

METHODE DE DETECTION AUTOMATIQUE DE CHANGEMENT DE SESSION DE RECHERCHE EN RECHERCHE D'INFORMATION PERSONNALISEE

Mémoire soutenu publiquement le 08/10/ 2019 devant le jury composé de :

Président : Mme Fellag Samia

Examinatrice : Mme Bourkache Ghenima

Encadré par : Mme Achmoukh Farida

Résumé

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.

La personnalisation est un processus qui change la fonctionnalité, l'interface, la teneur en information, ou l'aspect d'un système pour augmenter sa pertinence personnelle en intégrant le profil utilisateur.

La notion de session est une notion clef en recherche d'information personnalisée. Elle regroupe des interactions de recherche d'un utilisateur correspondante à un besoin en information spécifique.

Malgré des points de divergence, la majorité des définitions s'accordent sur le fait qu'une session permet de regrouper des requêtes soumises par un même utilisateur, liées à un même besoin en information.

Les sessions de recherche de distinguent par la durée, plutôt courte ou pouvant aller jusqu'à quelques heures mais dans tous les cas inférieure à une journée.

Selon la structure des sessions adoptée, les méthodes de détection automatique font appel à des caractéristiques des sessions et des ressources différentes. Ces méthodes peuvent ainsi exploiter la durée des sessions, le contenu lexical des requêtes et des sources de connaissance externes.

Dans le cadre de recherche d'information personnalisée, le profil utilisateur doit être intégré dans le processus de détection de changement de session de recherche. Différent travaux ont tenté de définir des méthodes de détection de changement de session de recherche.

Dans notre travail, nous proposons une implémentation d'une méthode de détection automatique de changement de session de recherche en recherche d'information personnalisée lié au changement de besoin en information de l'utilisateur.

Mots clés : Recherche d'information, Recherche d'information personnalisée, Centre d'intérêt, Profil d'utilisateur, Sessions de recherche.

Remerciements

Au terme de la rédaction de ce mémoire, nous estimons qu'il est important d'accorder quelques lignes de reconnaissances à toute personne ayant contribué de prêt ou de loin.
Tout d'abord, nous adressons notre plus profonde gratitude à notre promotrice Mme ACHMOUKH FARIDA, qui a toujours su orienté nos recherches, et pour tout le temps qu'elle nous a accordé.

Nous tenons à remercier également les membres de jury d'avoir accepté d'évaluer notre travail.

Dédicaces

A nos chers parents,
A nos frères et sœurs,
A nos familles,
A nos amis(es).

Sommaire

Liste des figures

Liste des tableaux

Introduction

Cadre général est objectifs

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.

La personnalisation est un processus qui change la fonctionnalité, l'interface, la teneur en information, ou l'aspect d'un système pour augmenter sa pertinence personnelle en intégrant le profil utilisateur.

La notion de session est une notion clef en recherche d'information personnalisée. Elle regroupe des interactions de recherche d'un utilisateur correspondante à un besoin en information spécifique.

Malgré des points de divergence, la majorité des définitions s'accordent sur le fait qu'une session permet de regrouper des requêtes soumises par un même utilisateur, liées à un même besoin en information.

Les sessions de recherche de distinguent par la durée, plutôt courte ou pouvant aller jusqu'à quelques heures mais dans tous les cas inférieure à une journée.

Selon la structure des sessions adoptée, les méthodes de détection automatique font appel à des caractéristiques des sessions et des ressources différentes. Ces méthodes peuvent ainsi exploiter la durée des sessions, le contenu lexical des requêtes et des sources de connaissance externes.

Dans le cadre de recherche d’information personnalisée, le profil utilisateur doit être intégré dans le processus de détection de changement de session de recherche. Différent travaux ont tenté de définir des méthodes de détection de changement de session de recherche.

Contribution

Dans notre travail, nous proposons une implémentation d'une méthode de détection automatique de changement de session de recherche en recherche d'information personnalisée lié au changement de besoin en information de l'utilisateur.

Organisation de la thèse

Notre travail est réparti sur 3 chapitres :

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

Chapitre 2 : L'accès personnalisé à l'information : Modélisation et évolution du profil utilisateur

Chapitre 3 : Evaluation expérimental délimitation des approches de session de recherche

Chapitre I : L’accès personnalisé à l’information : de la RI classique à la RI personnalisée

I.1. Introduction

La recherche d'information est un domaine qui permet l'acquisition, l'organisation, le stockage, la recherche et l'accès à l'information.

Son principale objectif est de relier le besoin en information d'un utilisateur exprimé par une requête avec un ensemble de documents en estiment leur pertinence par rapport a cette requête.

Les premiers modèles proposés en recherche d'information quantifient la pertinence d'un document seulement en fonction de la requête [Roberston et al. 1998].Or pour une même requête soumise par deux utilisateurs distincts et présentant des besoins différents, ces modèles retournent des résultats similaires.

De ce fait, les performances d'un SRI ne sont plus dépendantes de l'indexation des documents et de l'appariement document-requête uniquement, mais aussi sa capacité à modéliser l'utilisateur et son exploitation dans le processus de recherche d'information.

La première direction des travaux ayant apporté des solutions à cette problématique s'apportent à la recherche d'information adaptative. Son objectif consiste à améliorer le calcul du score de pertinence des documents grâce aux techniques de reformulations de requête [Rocchio 1971] et les techniques de filtrage d'information [Belkin et al. 1992]. Ces techniques ont ouvert de nombreuses perspectives, centralisant l'utilisateur au sein du processus de recherche d'information, faisant émerger un nouveau domaine, appelé recherche d'information orienté utilisateur.

Dans ce chapitre, nous présentons tout d'abord les concepts de base liés à la recherche d'information, en particulier les notions de document, de requête et de pertinence.la section 1.2 traite le processus de la recherche d'informations et ses différentes étapes, à savoir l'indexation, l'appariement et la reformulation d'un besoin en information. La section 1.3 passe en revue les principaux modèles de RI. La section 1.6 présente les principales causes d'émergence de la RI classique à la RI adaptative. Dans la section 1.7 nous introduisons la RI adaptative. La dernière section conclut le chapitre.

I.2. Les fondements de la recherche d'information

L'objectif principal de la recherche d'information est de fournir des techniques et des outils pour déterminer les informations pertinentes contenues dans un corpus en réponse aux besoins d'information d'un utilisateur qui est représentés à l'aide d'une requête.

Cette définition fait apparaître deux notions clés que nous introduisons dans ce qui suit :

Document et requête utilisateur.

I.2.1. Les concepts de base de la RI

I.2.1.1 Définition de la RI

La recherche d'information (RI) traite de la représentation, du stockage, de l'organisation et de l'accès à l'information.

La RI fournit donc les techniques et outils pour permettre de représenter, stocker, organiser, rechercher et retrouver, dans une masse documentaire existante, les documents contenant l'information qui répond au besoin informationnel exprimé par l'utilisateur sous forme de requête.

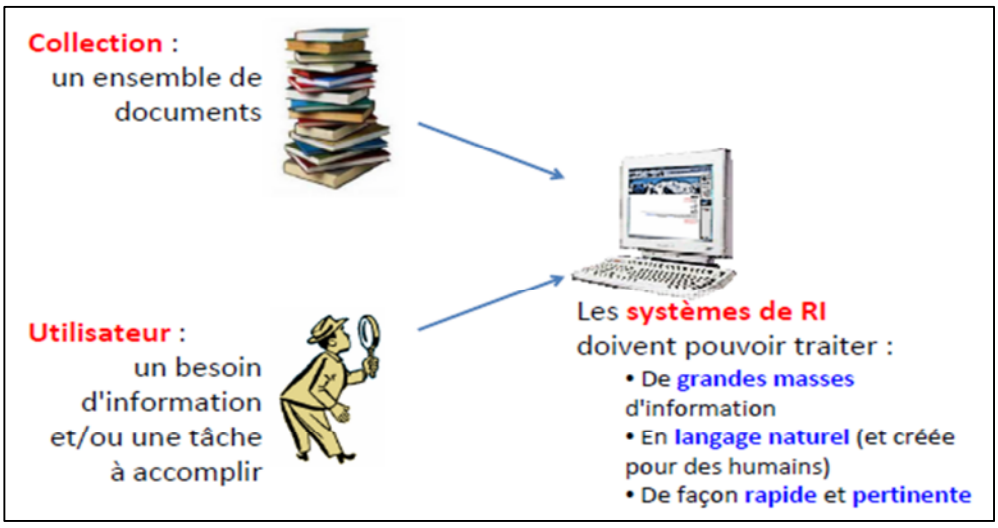


Figure I.1. Processus de la RI

Les systèmes de RI : Un système de recherche d'information (SRI) est un ensemble de programmes informatiques qui permet de retrouver, à partir d'une collection de **documents**, les documents **pertinents** pour une **requête** utilisateur.

Cette définition fait ressortir trois notions clés : document, requête, pertinence.

Document : Un document peut être un texte, un morceau de texte, une page

Web, une image, une vidéo, etc. On peut appeler document toute unité qui peut constituer une réponse à un besoin informationnel de l'utilisateur. Nous nous intéressons uniquement, dans ce travail, aux documents textuels. Dans la suite de ce mémoire, nous utilisons indifféremment les termes document ou information pour désigner l'utilité documentaire retournée en réponse à la requête de l'utilisateur.

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

On peut citer :

. Par une liste de mots clés : cas des systèmes SMART [149] et Okapi [143],

. En langage naturel : cas des systèmes SMART [149] et SPIRIT [61],

. En langage booléen : cas du système DIALOG [27],

. En langage graphique : cas du système NEURODOC [109].

La pertinence est une notion centrale en RI : Elle définit le degré de correspondance entre un document et une requête. Cette correspondance peut être considérée du point de vue de l'utilisateur (on parle alors de pertinence utilisateur), ou du point de vue système (on parle de pertinence système) :

- ✓ *La pertinence système* : c'est l'évaluation par le SRI, de l'adéquation entre des documents et une requête.
- ✓ *La pertinence utilisateur* : c'est l'évaluation par l'utilisateur, de la pertinence, vis-à-vis de son besoin en information, des documents retrouvés par le SRI.

I.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 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 [14] et schématiquement illustré par la figure 1.1.

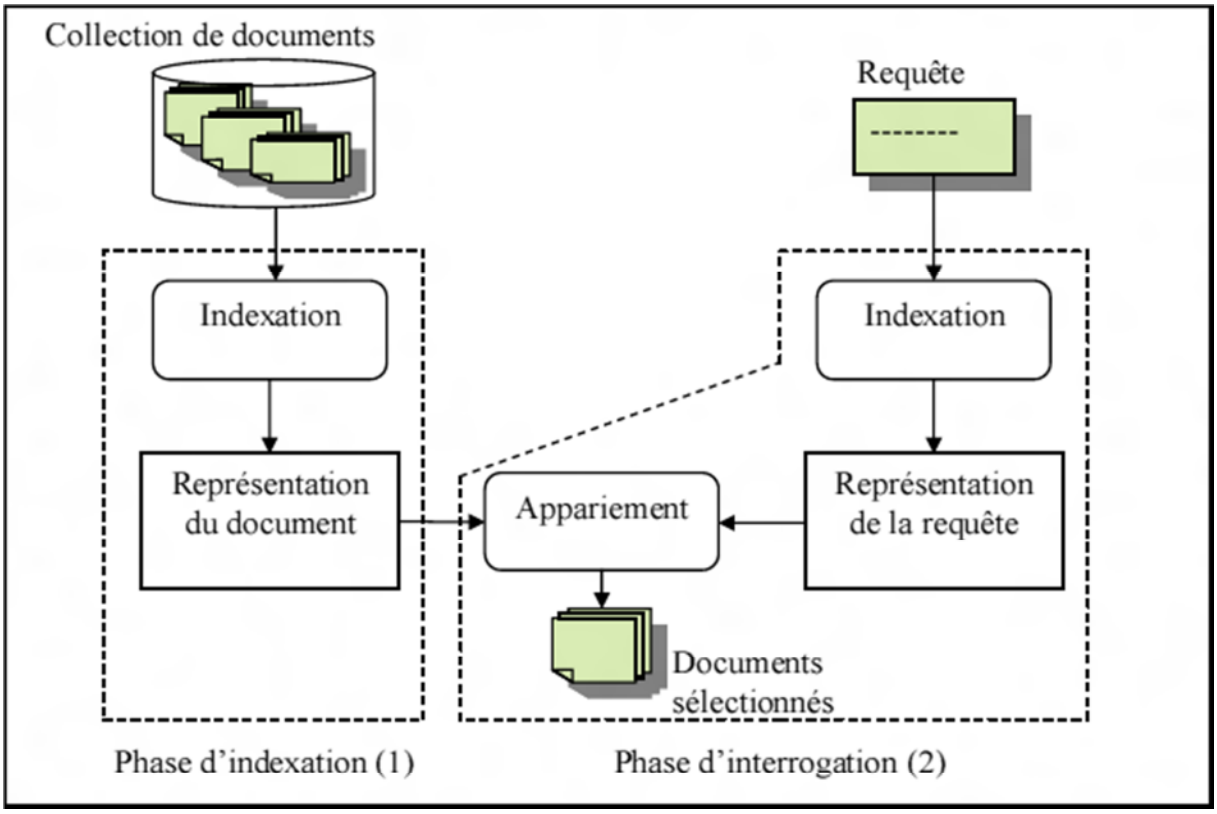


Figure I.2. Processus en U de la RI

Ce processus induit deux principales phases : indexation et appariement requête/document.

I.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.

Figure I-3 : Le processus de l'indexation automatique

D'une manière général, l'indexation automatique se fait en plusieurs étapes :

L'analyse lexical (Tokénisation en anglais) : consiste à découper le texte d'un document (ou d'une requête), en plusieurs unités lexicales représentant les termes d'index (Tokens) [Fox, 1992].

L'élimination des mots vides : vise à éliminer les mots non porteurs de sens ou mots vide. Ces mots peuvent être des mots outils tel que : les déterminations(Le, La…), des prépositions (sur, contre…), comme ils peuvent être les mots les plus fréquents par exemple : si un mot apparait dans plus de 80% des documents, alors il est jugé no utile pour la recherche.

L'élimination des mots peut se faire en utilisant une liste prédéfinie de mots vides (anti-dictionnaire) dite stopliste, ou en écartant les mots trop fréquents ou trop rares dans la collection.

La normalisation : permet de représenter les variantes morphologiques des termes, issus d'une même famille sous une forme normale. La normalisation ce base sur l'une des deux procédures :

La racinisation (troncature) **:** vise à supprimer l'affixe pour avoir des mots Sous forme tronquée, commune a toutes les variantes morphologiques.

Ex 1 : en anglais, la racinisation de « fishing' », « fished » , « fish » et « fisher » donne « fish ».

Ex 2 : cheval, chevaux, chevalier, chevalerie, chevaucher⇒« cheva » (mais pas « cavalier »)

La lemmatisation **(Stemming en anglais)** : permet l'obtention d'une forme canonique a partir d'un mot, les verbes sont transformés a l'infinitif, les normes et les adjectifs...Sont transformé en masculin singulier.

– Pour un verbe : sa forme à l'infinitif (sans les flexions)

Montrer, montreras, montraient → montrer.

– Pour un nom, adjectif, article, sa forme au masculin singulier

Vert, vertes, verts → vert.

La pondération

La pondération est une fonction fondamental en RI. Tous les modèles de recherche, excepté le model booléen, se basent sur la pondération des termes. L'idée de la pondération est d'affecter à chaque terme **t** d'un document **d** ou d'une requête **q,** un poids numérique sensé le caractériser dans le document ou la requête, les poids des termes de la requête et du document peuvent avoir des sémantique différentes, le poids est donc une mesure statistique de l'importance du terme dans le document (plus un terme est important dans un document, plus son poids dans ce document doit être élevé).

Parmi les mesures de pondération utilisé, nous avons la mesure $\mathbf{tf_{t,d}*idf_t}$

Notant : $\mathbf{tf_{t,d}}$ (term frequency) :La fréquence d'occurrence du terme **t** dans le document **d. C**ette mesure est proportionnelle à la fréquence du terme dans le document. Plus un terme est fréquent dans un document, plus un terme est fréquent dans un document, plus il est important dans la description de ce document.

$\mathbf{idf_t}$ (inverse of document Frequency) : La Fréquence documentaires **inverse** du terme **t**, c'est une 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.

Donc le poids $\mathbf{w_{t,d}= tf_{t,d}*idf_t}$. (I.1)

- 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.

I.2.2.2. L'appariement document-requête

Elle permet de mesurer la valeur de pertinence d'un document vis-à-vis d'une requête. LE SRI représente le document et la requête avec un même formalisme, puis il compare les deux représentations, afin d'obtenir un résultat qui détermine le degré de ressemblance du document avec la requête **[Hammache, 2011].**

Il existe deux types d'appariements :

Appariement exact : les documents retournés respectent exactement la requête spécifié avec des critères précis, ces documents sont tries.

Appariement approché : les documents retournés répondent a tout ou a une partie de la requête, ces documents sont tries selon un ordre de mesure. Cet ordre reflète le degré de pertinence document-requête.

I.3. Taxonomie des modèles de RI

Les travaux de recherche dans le domaine de la RI ont conduit à la proposition de nombreux modèles **[R. Baeza-Yates and R. A. Ribeiro-Neto, 99].**

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.

I.3.1. Le modèle booléen

Les premiers SRI développés sont basés sur le modèle booléen, même aujourd'hui beaucoup de systèmes commerciaux (moteurs de recherche) utilisent le modèle booléen. Cela est dû à la simplicité et à la rapidité de sa mise en œuvre.

Le modèle booléen est basé sur la théorie des ensembles et l'algèbre de Boole. Dans ce modèle, un document **d** est représenté par un ensemble de mots-clés (termes) ou encore un vecteur booléen.

La requête **q** de l'utilisateur est représentée par une expression logique, composée de termes reliés par des opérateurs logiques : ET (∧), OU (∨) et SAUF (¬).

L'appariement (RSV) entre une requête et un document est un appariement exact, autrement dit si un document implique au sens logique la requête alors le document est pertinent. Sinon, il est considéré non pertinent. La correspondance entre document et requête est déterminée comme suit :

$$RSV\ (q, d) = \{1,0\}. \qquad (1.2)$$

Le modèle booléen a des avantages qui sont présentés par :

Le modèle de recherche booléen est reconnu pour sa force pour faire une recherche très restrictive et obtenir, pour un utilisateur expérimenté, une information exacte et spécifique.

La simplicité du modèle le rend aisément compréhensible pour un utilisateur.

L'efficacité du modèle est due aux spécialistes qui ont explorés le corpus avec une bonne connaissance du vocabulaire.

La formulation des requêtes devient vite laborieuse quand la requête se fait précise (donc longue).

Ce modèle n'a pas seulement d'avantages, il a aussi des inconvénients qui sont les suivants :

L'inconvénient majeur de ce modèle comme schématisé dans la Figure I.3, est que les documents pertinents dont la représentation ne correspond qu'approximativement à la requête ne sont pas sélectionnés.

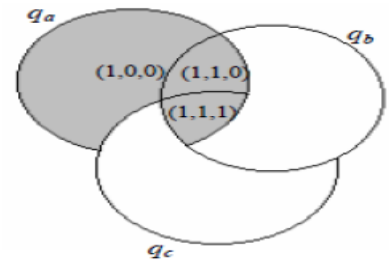


Figure I.4 : Les 3 composants conjonctifs pour la requête :

$q = q_a \wedge (q_b \vee \neg q_c)$

Tous les termes ont la même importance et il est incapable de trier les documents pertinents.

L'impossibilité de rendre compte d'une correspondance partielle d'un document à une requête.

La pondération binaire des termes du vocabulaire limite la pertinence des résultats et ne permet pas de les ordonner.

Exemple illustratif de l'inconvénient du model booléen :

d1: Le loup est dans la bergerie.

d2: Le loup et les trois petits cochons.

d3: Les moutons sont dans la bergerie.

d4: Spider Cochon, Spider Cochon, il peut marcher au plafond.

d5: Un loup a mangé un mouton, les autres loups sont restés dans la bergerie.

d6: Il y a trois moutons dans le pré, et un mouton dans la gueule du loup.

d7: Le cochon est à 12 le Kg, le mouton à 10 E/Kg.

d8: Les trois petits loups et le grand méchant cochon.

Et ainsi de suite. Supposons que l'on recherche tous les documents parlant de loups, de moutons mais pas de bergerie (c'est le besoin). Une solution simple consiste à parcourir tous les documents et à tester la présence des mots-clés. Ce n'est pas très satisfaisant car: Commençons par montrer une matrice d'incidence avec les documents en ligne. On se limite au vocabulaire suivant: {"loup", "mouton", "cochon", "bergerie", "pré", "gueule"}.

La matrice d'incidence

	Loup	**Mouton**	**Cochon**	**bergerie**	**pré**	**Gueule**
d1	1	0	0	1	0	0
d2	1	0	1	0	0	0
d3	0	1	0	1	0	0
d4	0	0	1	0	0	0
d5	1	1	0	1	0	0
d6	1	1	0	0	1	1
d7	0	1	1	0	0	0
d8	1	0	1	0	0	0

Tableau I.1 matrice d'incidence

Cette structure est parfois utilisée dans les bases de données sous le nom d'index bitmap. Elle permet de répondre à la recherche de la manière suivante. On prend les vecteurs d'incidence de chaque terme contenu dans la requête, soit les colonnes dans notre représentation.

Loup: 11001101

Mouton: 00101110

Bergerie: 01010111

On fait un **et** (logique) sur les vecteurs de Loup et Mouton et on obtient 00001100. Puis on fait un **et** (logique) du résultat avec le complément du vecteur de Bergerie (01010111) et on obtient 00000100, d'où on déduit que la réponse est limitée au document **d6**.

I.3.2. Le modèle vectoriel

Le modèle vectoriel de base a été introduit par Salton concrétisé dans le cadre du système SMART. Ce modèle se base sur une formalisation géométrique.

En effet, les documents et les requêtes sont représentés dans un même espace, défini par un ensemble de dimensions, chaque dimension représente un terme d'indexation.

Les requêtes et les documents sont alors représentés par des vecteurs, dont les composantes représentent le poids du terme d'indexation considéré dans le document (la requête).

Formellement, si on a un Espace T de termes d'indexation de dimension **n, T= {t1, t2, t3...t_n}.** Un document d_i est Représenté par un vecteur **d_i (w_{i1}, w_{i2}, w_{i3}, w_{i4},…w_{in})** Une requête **q** par un vecteur **q (w_{q1}, w_{q2}, w_{q3}, w_{q4},…w_{qn}).**

Où **w_{ij}** (resp. w_{qj}) représente le poids du terme t_j dans le document **d_i** (respectivement dans la requête q).

Le modèle vectoriel offre des moyens pour la prise en compte du poids de terme dans le document. Dans la littérature, plusieurs schémas de pondération ont été proposés. La majorité de ces schémas prennent en compte la pondération locale et la pondération globale.

La pondération locale permet de mesurer l'importance du terme dans le document. Elle prend en compte les informations locales du terme qui ne dépendent que du document. Elle correspond en général à une fonction de la fréquence d'occurrence du terme dans le document

(Noté tf pour term frequency), exprimée ainsi :

$$tf = 1 + (\log(freq(t, d)) \text{ si } freq(t, d) \neq 0 \qquad (I.3)$$

Où freq (t_i, d_j) est la fréquence du terme t_i dans le document d_j.

Quant à la *pondération globale,* elle prend en compte les informations concernant le terme dans la collection. Un poids plus important doit être assigné aux termes qui apparaissent moins fréquemment dans la collection. Car les termes qui apparaissent dans de nombreux documents de la collection ne permettent pas de distinguer les documents pertinents des documents non pertinents (i.e. peu utile pour la discrimination).

Un facteur de pondération globale est alors introduit. Ce facteur nommé idf (inverted document frequency), dépend d'une manière inverse de la fréquence en document du terme et exprimé comme suit :

$$idf = \log\left(\frac{N}{n_i}\right) \qquad (I.4)$$

Où $\mathbf{n_i}$ est la fréquence en document du terme considéré, et **N** est le nombre total de documents dans la collection.

Les fonctions de pondération combinant la pondération locale et globale sont référencées sous le nom de la mesure **tf*idf**.

Cette mesure donne une bonne approximation de l'importance du terme dans les collections de documents de taille homogène. Cependant, un facteur important est ignoré, la taille du document. En effet, la mesure (**tf*idf**) ainsi définie favorise les documents longs, car ils ont tendance à répéter le même terme, ce qui accroit leur fréquence, par conséquent augmentent la similarité de ces documents vis-à-vis de la requête.

Pour remédier à ce problème, des travaux ont proposé d'intégrer la taille du document dans les formules de pondération, comme facteur de normalisation.

L'appariement document-requête dans le modèle vectoriel, consiste à trouver les vecteurs documents qui s'approchent le plus de vecteur de la requête. Cet appariement est obtenu par

L'évaluation de la distance entre les deux vecteurs. Plusieurs mesures de similarité ont été définies dont les plus courantes sont décrites dans le tableau I.2 ci-dessous.

Mesures	Formules
Le produit scalaire	$RSV(q,d_i) = \sum_{j=1}^{\|T\|} w_{qj}.w_{ij}$
La mesure de cosinus	$RSV(q,d_i) = \frac{q.d_i}{\|q\|.\|d_i\|} = \frac{\sum_{j=1}^{\|T\|} w_{qj}.w_{ij}}{\sqrt{\sum_{j=1}^{\|T\|} {w_{qj}}^2 \sum_{j=1}^{\|T\|} {w_{ij}}^2}}$
La mesure de Dice	$RSV(q,d_i) = \frac{2 \times \sum_{j=1}^{\|T\|} w_{qj}.w_{ij}}{\sum_{j=1}^{\|T\|} {w_{qj}}^2 + \sum_{j=1}^{\|T\|} {w_{ij}}^2}$
La mesure de Jaccard	$RSV(q,d_i) = \frac{\sum_{j=1}^{\|T\|} w_{qj}.w_{ij}}{\sum_{j=1}^{\|T\|} {w_{qj}}^2 + \sum_{j=1}^{\|T\|} {w_{ij}}^2 - \sum_{j=1}^{\|T\|} w_{qj}.w_{ij}}$

Tableau I.2 : Les mesures de similarité utilisées dans le modèle vectoriel

Le modèle vectoriel a des avantages qui sont présentés par :

- ✓ Le modèle vectoriel est relativement simple à appréhender (algèbre linéaire) et est facile à implémenter.
- ✓ Il permet de retrouver assez efficacement des documents dans un corpus non structuré et cela dépend de la qualité de la représentation.

Ce modèle n'a pas seulement d'avantages, il a aussi des inconvénients qui sont les suivants :

- ✓ La représentation vectorielle permet une mise en correspondance des documents avec une requête imparfaite.
- ✓ Il comporte également plusieurs limitations qui furent, pour certaines, corrigées par des affinements du modèle.

L'indépendance des termes représentatifs supposée par le modèle.

Dans un texte l'ordre des mots, les synonymes, la morphologie des contenus ne sont pas pris en compte.

I.3.3. Le modèle probabiliste

Le modèle de recherche probabiliste utilise un modèle mathématique fondé sur la théorie de la probabilité. Le processus de recherche se traduit par calcul de proche en proche, du degré ou probabilité de pertinence d'un document relativement à une requête. Pour ce faire, le processus de décision complète le procédé d'indexation probabiliste en utilisant deux probabilités conditionnelles :

P (wji/pert) : Probabilité que le terme ti occurre dans le document Dj sachant que ce dernier est pertinent pour la requête.

P (wji/Nonpert) : Probabilité que le terme ti de poids dji occurre dans le document Dj sachant que ce dernier n'est pas pertinent pour la requête.

Si on suppose l'indépendance des variables documents « pertinents » et « non

pertinents », la fonction de recherche peut être obtenue en utilisant la formule de Bayes.

Soit : Dj (t1, t2,…, tN) Où :

$t_i = 1$ si ti indexe le document Dj,

0 sinon.

$$P(\text{pert}, Dj) = (p(Dj/\text{pert}) * p(\text{pert}))/p(Dj) \quad \text{(I.5)}$$

et

$$P(\text{Nonpert}, Dj) = (p(Dj/\text{Nonpert}) * p(\text{Nonpert}))/p(Dj) \quad \text{(I.6)}$$

Avec :

P (pert/Dj) : est la probabilité de pertinence du document Dj sachant sa description.

$$P(Dj) = p(Dj/\text{pert}) * p(\text{pert}) + p(Dj/\text{Nonpert}) * p(\text{Nonpert}) \quad \text{(I.7)}$$

p (Dj /pert) (respectivement P (Dj /Nonpert) est la probabilité d'observer le document Dj sachant qu'il est pertinent (respectivement non pertinent) .

Le modèle probabiliste a des avantages qui sont présentés par :

- ✓ Il a une base théorique saine et il est indépendant du domaine d'application.

Pour des raisons de simplicité, l'hypothèse de l'indépendance des termes est utilisée en pratique pour implémenter ces modèles.

I.4. Evaluation d'un système de recherche d'information

La qualité d'un système doit être mesurée en comparant les réponses du système avec les réponses idéales que l'utilisateur espère recevoir, plus les réponses du système correspond à celles que l'utilisateur espère, mieux est le système.

La comparaison des réponses d'un système pour une requête avec les réponses idéales nous permet d'évaluer les deux métriques suivantes:

I.4.1. Précision

La précision mesure la proportion de documents pertinents retrouvés parmi tous les documents retrouvés par le système.

$$\text{Précision i} = \frac{\text{nb de document correctement attribué a la classe i}}{\text{nb de documents attribués a la class i}} \qquad \text{(I.8)}$$

I.4.2. Rappel

Le rappel mesure la proportion de documents pertinents retrouvés parmi tous les documents pertinents dans la base.

$$\text{rappel i} = \frac{\text{nb de document correctement attribué a la classe i}}{\text{nb de documents appartenant a la classe i}} \qquad \text{(I.9)}$$

Un système idéal obtient des taux de précision et de rappel proches de 1. Un système qui obtiendrait une précision de 1 et un rappel de 1 signifie qu'il retrouve tous les documents pertinents, et uniquement les documents pertinents : la pertinence système et la pertinence utilisateur seraient confondues.

Pour faire une évaluation de la qualité d'un système de recherche d'information, on utilise habituellement une collection de test. Une collection de test contient un corpus de documents, un ensemble de requêtes, et la liste des documents pertinents pour chaque requête.

Cette liste de réponses idéales est établie par des experts ayant une grande connaissance du corpus et du domaine des documents. Pour chaque requête, on établit alors une courbe de la précision en fonction du rappel. La moyenne de ces courbes constitue un profil visuel de la qualité d'un système.

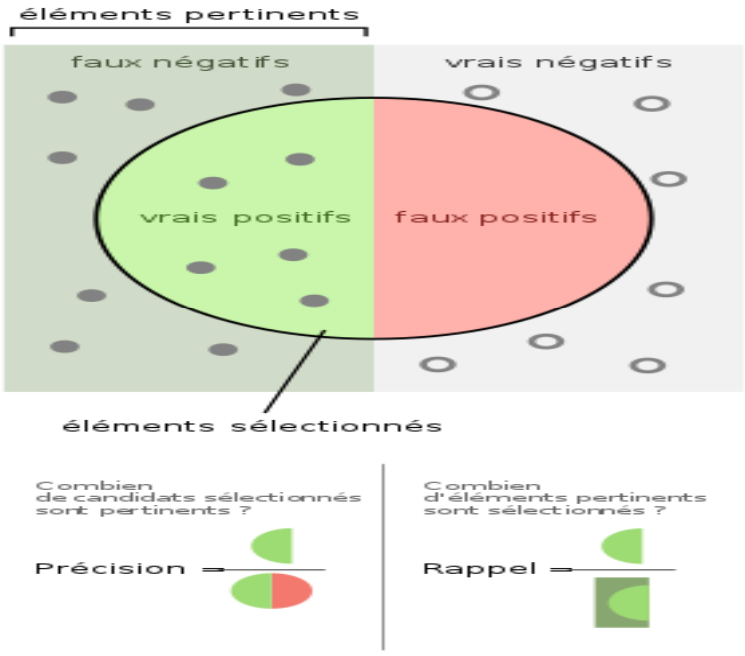


Figure 1.5 : Différence entre Précision et Rappel

I.4.3 La courbe de Précision-Rappel

Dans le cas d'un système idéal, le taux de précision est égal au taux de rappel, c'est-à-dire que, tous les documents pertinents dans ce cas, et que ceux-ci, sont sélectionnés. On aurait donc une droite.

En pratique, la courbe de Précision-Rappel à l'allure générale de la Figure suivante. Pour ce faire on procède comme suit :

Pour i=1, 2,3,…#de_documents_dans_la_base faire :

Evaluer la précision et le rappel pour les i premiers documents dans la liste des réponses du système.

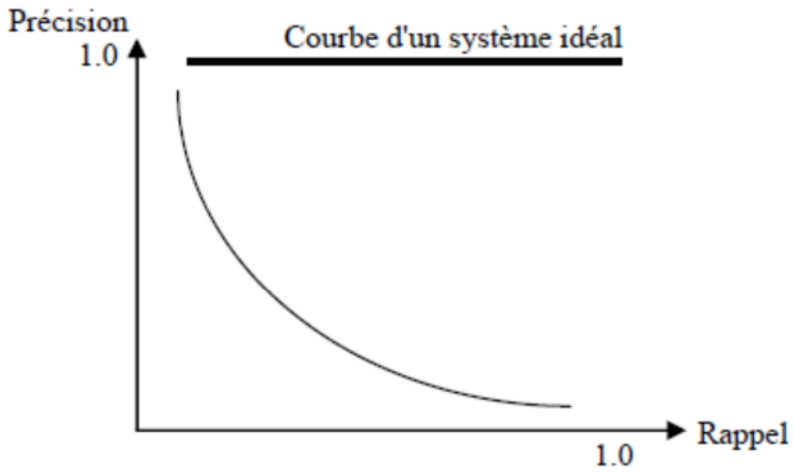


Figure1.6 : Allure d'une courbe de rappel-précision.

I.5. Les Campagnes d'évaluations

Les campagnes d'évaluation en RI permettent d'évaluer sur des collections différentes plusieurs SRI, afin de valider les différents modèles mis en œuvre, et comparer les systèmes.

Les objectifs essentiels des campagnes sont les suivants :

- Encourager la RI sur de grandes collections fermées.

- Développer la communication entre l'industrie, l'académie et le gouvernement en mettant en place un forum ouvert pour faciliter les échanges d'idées sur la recherche.

- Augmenter la vitesse de transfert de la technologie du laboratoire de recherche aux enseignes commerciales.

- Rendre disponible et accessible des techniques d'évaluations appropriées pour les industriels et les académiciens.

Chaque campagne est constituée d'un certain nombre de tâches fournissant des résultats, et un protocole d'évaluation pour chaque tâche.

Les campagnes de TREC (Text REtrieval Conference) qui ont vu le jour en 1992 avec 25 participants issus du monde académique et industriel sont devenues la référence en ce qui concerne l'évaluation des systèmes de recherche d'information.

I.6. De la RI classique à la RI adaptive

Le principe fondamental commun à tous les modèles classiques de RI suppose que les documents sélectionnés doivent contenir les mêmes mots (voir une partie) que ceux formulés par l'utilisateur et que la requête représente ce besoin en information.

Dans le cas du modèle booléen par exemple, le document sélectionné doit contenir tous les mots (cas conjonctif) ou une partie des mots (cas disjonctif) de la requête.

Dans le cas du modèle vectoriel, plus un document partagé des mots avec la requête et dans la même proportion de poids, plus grande n'est sa similarité avec la requête. Ainsi, l'efficacité du procédé de sélection naïve de ces modèles, repose principalement sur l'efficacité et la qualité des mécanismes d'indexation et d'appariement **[N. Belkin and W. Croft,1992].**

Lors de l'appariement requête/document, seuls les documents qui sont les plus proches sémantiquement du besoin de l'utilisateur sont sélectionnés. De ce fait, plus les termes d'indexation ne sont représentatifs du contenu sémantique des documents et de la requête, plus la pertinence des documents sélectionnés est améliorée. Néanmoins, dans la pratique la majorité des requêtes exprimées par les utilisateurs sont courtes et ambiguës **[C. Bradford**

and I. Marshall, 99], ce qui donne des spécifications inachevées sur leur besoin en informations.

I.7. La RI adaptive

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

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

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

I.7.1. Reformulation de requête

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

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

I.7.1.1 Reformulation automatique de requête

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

Plus précisément, au niveau des ressources linguistiques, le but est d'utiliser un

Vocabulaire contrôlé issu de ressources externes. Il s'agit principalement de chercher des associations inter-termes extraites à partir des ontologies linguistiques (tel que WordNet [**G. Miller, 95**]), ou à partir de thésaurus [E. Voorhees, 94 ; G. Brajnik, 96].

Les thésaurus construits manuellement sont un moyen efficace pour l'expansion de requête.

I.7.1.2 Reformulation interactive de requête

A la différence de la reformulation automatique, l'approche interactive (ou par réinjection de pertinence et/ou non-pertinence) exploite uniquement un sous-ensemble de documents sélectionnés parmi les premiers résultats obtenus de l'exécution de la requête initiale.

Son principe fondamental est d'utiliser cette requête pour amorcer la recherche, puis modifier celle-ci à partir des jugements de pertinence et/ou de non pertinence de l'utilisateur, soit pour repondérer les termes de la requête initiale [S. Robertson et all 00 ; L. Tamine et all, 03], soit pour y ajouter (resp. supprimer) d'autre termes contenus dans les documents pertinents (resp. non pertinents) [J. Rocchio, 71].

La nouvelle requête ainsi obtenue à chaque itération de feedback, permet de corriger la direction de la recherche dans le sens des documents pertinents.

Plusieurs techniques ont été introduites dans différents modèles de recherche [J. Rocchio, 71; K. Kwok, 89 ; M. Boughanem and C. Soulé-Dupuy, 97; M. Boughanem et all, 99; S. Robertson et all, 00; L. Tamine, 00; M. Boughanem et all 2002], notamment dans les modèles vectoriel et probabiliste, décrits ci-après.

I.7.2 Adaptation du contenu documentaire

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

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

Un réseau de neurone formel est construit à partir des représentations initiales des documents et de la requête. Le mécanisme de recherche d'information est fondé sur le principe de propagation de valeurs depuis les neurones descriptifs de la requête vers ceux des documents, à travers les connexions du réseau.

Les résultats sont présentés à l'utilisateur selon le niveau d'activation des neurones documents. Le modèle connexionniste est connu pour sa capacité d'apprentissage, ce qui permet aux SRI de devenir adaptatifs.

Plusieurs modèles basés sur le principe des réseaux de neurones sont utilisés en RI [R. Belew, 89; K. Kwok ,89; M. Boughanem and C. Soulé-Dupuy, 97 ; F. Crestani, 93]. Cependant, il n'existe pas de représentation unique d'un réseau de neurones pour la recherche d'information, c'est au constructeur du modèle de le définir, et ce en identifiant les éléments suivants :

- Les différentes couches1 du réseau (couche d'entrée, de sortie, intermédiaires, etc.) ;
- les neurones de chaque couche,
- la fonction d'entrée de chaque neurone,
- la fonction de sortie de chaque neurone,
- les liens entre les neurones et leurs poids associés.

Les travaux de [R. Belew, 89] sont parmi les premiers a avoir abordé l'approche connexionniste en

RI. AIR (Adaptive Information Retrieval), le système proposé dédié à la recherche dans le domaine bibliographique, est construit autour d'un réseau à trois couches : auteurs, termes et documents. Les liens entre les termes et les documents sont initialement pondérés par id f (le nombre de fois qu'un terme apparaît dans un document).

Le système utilise les jugements des utilisateurs pour modifier ces liens dans le but d'arriver à une représentation consensuelle des termes dans les documents partagés par les utilisateurs.

Les modèles à couches, les plus performants de ces dernières années, sont ceux proposés par Kwok [K. Kwok , 89] dans le système de recherche d'information PIRCS et par Boughanem [M. Boughanem, 92; M. Boughanem and C. Soulé-Dupuy, 97] dans le système de recherche d'information MERCURE (ModèlE de Réseau Connexionniste poUr la REcherche d'information) :

Le modèle PIRCS est un réseau à couches interconnectées dans le sens requête(Q)-termes(T)-documents(D) [K. Kwok , 89]. Les connexions sont bidirectionnelles et asymétriques. L'approche de Kwok est fondée sur l'idée que les requêtes et documents sont similaires. Sur cette base, elle reprend des éléments du modèle probabiliste pour classer les neurones documents, répondant à une requête selon la probabilité, donnée par la formule suivante :

$$Wd = Wqd + Wdq \qquad (1.10)$$

Où :

Wqd : est la probabilité que la requête q soit pertinente pour le document d, et Wdq la probabilité pour que le document d soit pertinent pour la requête q.

I.8. Bilan sur la RI adaptative : facteurs d'émergence de la RI personnalisée

Les travaux en RI adaptative ont certes apporté des solutions en particulier au défaut d'appariement requête-document qui ont conduit à l'amélioration des performances du processus de recherche d'information [D. Harman, 92].

Cependant une analyse fine des travaux dans ce domaine montre que ces performances dépendent de nombreux facteurs a priori non contrôlés par le processus de réécriture adaptative de la requête.

Ces facteurs, principalement liés aux approches de reformulation de requête, qui sont ainsi problématiques, peuvent être catégorisés selon trois principales dimensions : l'utilisateur, l'information portée par la requête et/ou document et l'interaction entre l'utilisateur et le SRI. Nous discutons dans ce qui suit chacune de ces dimensions [L. Tamine and S. Calabretto, 08].

I.8.1. La dimension utilisateur

(a) l'expression initiale du besoin en information de l'utilisateur (ce qu'il ne sait pas du sujet de la requête) dépend de ses centres d'intérêt (ce qu'il sait déjà du sujet de la recherche) et de ses buts [P. Ingwersen, 96]. Cependant, ces éléments ne se déclinent pas dans le processus de réécriture de la requête initiale.

(b) [I. Hsieh-Yee, 93] montre qu'il existe une corrélation positive entre familiarité de l'utilisateur avec le sujet de la requête et les performances de la stratégie de réinjection de la pertinence.

De plus, le niveau d'expertise de l'utilisateur [I. Ruthven and M. Lalmas, 03; R. W. White et all 03] a un impact sur les performances de recherche. En ce sens que des utilisateurs expérimentés effectuent de meilleurs choix quant à la qualité des documents et termes utilisés pour la réécriture de la requête, relativement à des utilisateurs novices.

(c) [R. Fidel, 91] montre que la discipline professionnelle de l'utilisateur n'est pas sans impact dans la perception de l'information et donc de la pertinence. Ceci influe directement sur les performances de recherche.

(d) la nature (utilité, intérêt, préférence) et valeur du jugement de pertinence de l'utilisateur (peu pertinent, très pertinent, assez pertinent etc.) dépend de nombreux facteurs :

(1) de ses centres d'intérêt et ses buts [P. Vakkari, 00].

(2) de l'objet de la requête (Ce qui est attendu à travers une requête : service, information, page de référence)[L. B. Lorigo, H. Pan, T. Hembrooke, 06; I. Kang and G. Kim, 04],

(3) de la complexité de la tâche de recherche qui est déterminée par la quantité d'information que doit traiter l'utilisateur pour atteindre l'information pertinente [P. Vakkari, 01].

Cependant, la RI adaptative exploite des jugements de pertinence binaire supposés ne dépendre que du contenu des documents.

I.8.2. La dimension information

(a) le volume important d'information accessible engendre incontestablement une diversité importante du vocabulaire. Par conséquent, les algorithmes d'ordonnancement des termes d'expansion de requêtes en fonction de leur corrélation au sujet de la requête, sont peu performants [W. Croft, R. Cook, 95].

(b) les documents du Web contiennent de nombreuses informations non directement liées au sujet du document telles que les liens de navigation, les informations ou images publicitaires etc. Ces informations, même extraites des documents les mieux classés à l'issue d'une recherche initiale, engendrent du bruit lors d'un processus de réécriture de requête [S. Yu, D. Cai, J. Wen, and W. Ma, 03].

(c) les stratégies classiques de réinjection de pertinence sont peu capables de rappeler des documents traitant de différents sujets auxiliaires associés à un sujet fédérateur véhiculé par la requête [C. Zhai and J. Cohen, 03]. Le même problème est posé avec des documents traitant de nombreux sujets à la fois tels que les journaux [S. Yu, D. Cai, J. Wen, and W. Ma, 03].

I.8.3. La dimension interaction

(a) Les processus de réinjection de la pertinence induisent une interaction qui est à l'origine d'une surcharge cognitive pour l'utilisateur. La valeur ajoutée de ces interactions dépend du degré de participation de l'utilisateur.

De plus, des études ont montré [N. Belkinet all, 01; R. W. White et all, 03] que les utilisateurs n'utilisent pas forcément l'ensemble des possibilités offertes par le système quant à l'enrichissement de la requête et ce, pour une raison majeure : les utilisateurs n'en cernent pas le principe et le lien avec l'opération de sélection de l'information pertinente.

(b) la forme de présentation des documents (Titre, résumé, texte plein) exploités pour la réinjection de la pertinence a un impact non négligeable sur le jugement de l'utilisateur [J. Janes, 91].

(c) l'utilité de la réinjection de pertinence est plus déterminante aux dernières itérations d'un processus de recherche d'information adaptative [X. Shen, B. Tan, and C. Zhai, 05].

Ce bilan montre globalement que les stratégies de RI adaptative ne sont pas garantes sur l'uniformité de la qualité des résultats d'un SRI dans des conditions d'utilisation différentes. Il en ressort que les éléments clés à intégrer dans de telles stratégies dans le but d'en améliorer les performances, sont dépendants les uns des autres, liés cependant à différentes dimensions.

De ce fait, le développement de services d'accès délivrant l'information pertinente de manière personnelle en fonction des caractères spécifiques de l'utilisateur et adaptant les résultats de recherche en fonction des préférences et contexte de l'utilisateur devient une nécessité absolue. L'ensemble de ces éléments constitue l'ensemble des facteurs précurseurs ayant déterminé les directions d'investigation pour le développement de la troisième génération des SRI.

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

I.9. Conclusion

Ce chapitre a porté essentiellement sur les notions de base de la Ri classique ainsi que l’émergence de la RI orienté utilisateur. Nous passé en revue le processus de RI et les étapes qui le composent, où nous avons montré les différents modèles permettent de mesurer la similitude requête-document, afin de déterminer les documents pertinents.

Dans un second temps nous avons pris en compte le profil utilisateur pour la personnalisation de l’information.

Nous présentons dans le prochain chapitre les principaux concepts et techniques liés à la recherche d’information personnalisée.

Chapitre II : L’accès personnalisé à l’information : Modélisation et évolution du profil utilisateur

II.1. Introduction

L'objectif de la personnalisation de l'information consiste à délivrer des informations pertinentes, adaptés, précis et aux préférences de l'utilisateur **[Pitkow et al.2002]**. La pluparts des approches modélisent l'utilisateur dans une structure informationnelle appelée profil [Bennet et al. 2002]. Les premières approches le représentent comme une ou plusieurs vecteurs définis dans un espace de terme d'indexation **[Gowan 2003],** puis d'autres l'ont organisé comme un modèle structuré de dimension prédéfinies **[Kostadinov 2007].**La construction du profil de l'utilisateur peut être explicite fourni directement par l'utilisateur **[Maghoul et al. 2005]** , ou implicite à partir des documents consultés et du comportement de l'utilisateur **[Sasel, 2010].**

Cependant, la personnalisation de l'information engendre le problème de l'évolution du profil de l'utilisateur au cœur du temps. Dans la plupart des systèmes de recherche personnalisée, l'évolution du profil est exprimée par l'ajout de nouvelles informations, en adaptant son contenu aux variantes des besoins en information de l'utilisateur **[Bericha-Bohe et al.2007].**

Nous présentons dans ce chapitre la personnalisation de la recherche d'information et la modélisation du profil utilisateur ainsi que son évolution au cours du temps. Nous nous intéressons dans la section 2.2 à la notion de contexte et de profil utilisateur. La section 2.3 traite les différentes approches et techniques de modélisation du profil de l'utilisateur, à savoir sa représentation, sa construction et son évolution.

Dans la section 2.4 nous abordons les méthodes de détection automatique de session de recherche. L'architecture fonctionnelle d'un système de recherche personnalisée est présentée dans la section 2.5 avec les modèles d'accès à l'information. La section 2.6 traite les protocoles d'évaluation des systèmes d'accès personnalisé à l'information. La dernière section conclut le chapitre.

II.2. La notion de profil utilisateur

La notion de profil utilisateur a été largement abordée dans le domaine du *user modeling*. Depuis le début des années 70, les recherches menées dans ce domaine se sont principalement focalisées sur la possibilité de définir des approches de modélisation de l'utilisateur dans le contexte de différentes applications **[W. Pohl 1997].** 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.

Plusieurs définitions du profil ont été abordées dans la littérature en RI personnalisée.

On distingue le profil cognitif, le profil qualitatif et le profil multidimensionnel.

- Le profil qualitatif est lié aux préférences de recherche de l'utilisateur

Quant la qualité de l'information restituée par le système (fraîcheur, crédibilité des sources d'information, cohérence, etc.…) ces critères concernent le contexte du document qualitatif

Dans certaines études, le profil couvre en plus des centres d'intérêts et des préférences de l'utilisateur, des caractéristiques de l'environnement et du système, définissant ainsi un profil multidimensionnel **[G. Koutrika & al. 2005] [B.Tan & al. 2006].**

- Un profil multidimensionnel se rapproche de la notion du contexte multidimensionnel et qui couvre toutes les dimensions possibles impliquées dans l'interaction de l'utilisateur avec le système.

On appel 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 **[W. N. Zemirli & al. 2005], [M. Bouzeghoub & al. 2005].**

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

le centre d'intérêt et les préférences de l'utilisateur peuvent légitimement évolué **[A. Kobsa &al. 2007].**

II.3. Architecture fonctionnelle d'un SRIP :

Le but fondamental d'un SRI personnalisée est de satisfaire les besoins en information de l'utilisateur en intégrant son profil dans la chaîne d'accès à l'information.

Le *SRIP* ne se limite pas seulement à modéliser les caractéristiques des utilisateurs en des profils.

Il doit être capable de déduire à partir de ces profils, l'intention de l'utilisateur lorsqu'il effectue sa recherche, en d'autres termes son **contexte de recherche**, et de détecter **l'évolution des profils** de manière dynamique.

Le système doit donc inclure :

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

La diversité des systèmes de personnalisation rend la définition d'une architecture fonctionnelle et formelle d'accès personnalisé à l'information difficilement généralisable. Néanmoins, nous tentons dans ce paragraphe de dégager une architecture standard pour un système d'accès personnalisé à l'information, présentée par la figure 2.1.

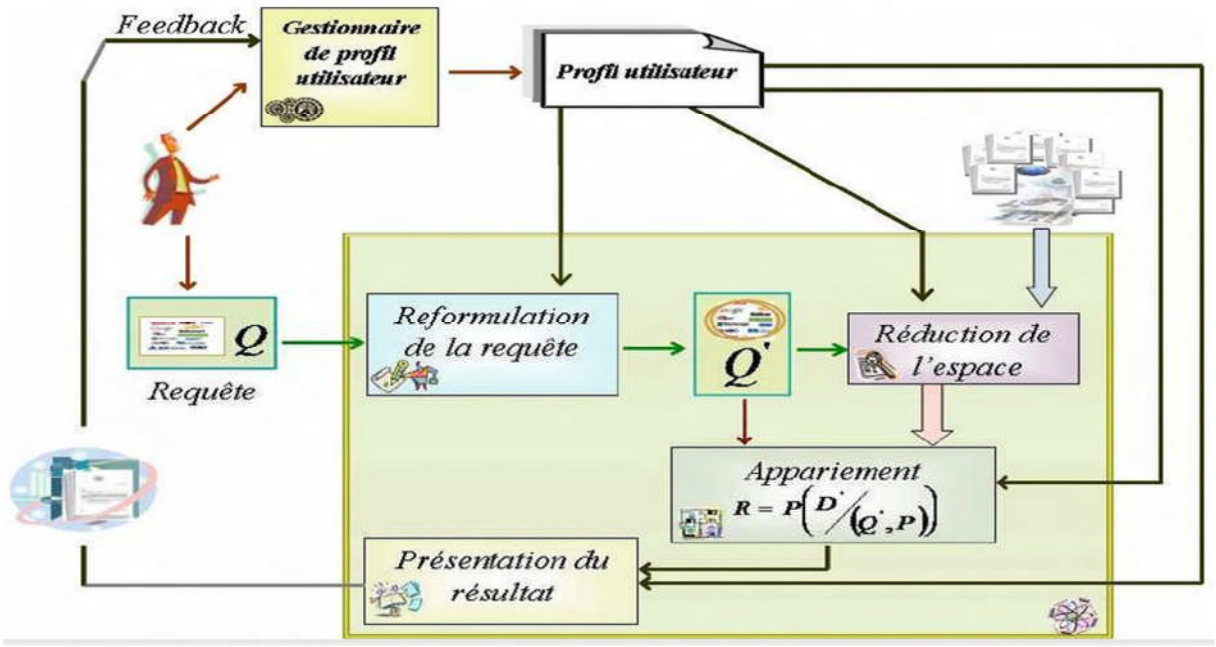


Figure II.1 : Processus général d'accès personnalisé à l'information.

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

1. *Un gestionnaire de profil* pour représenter, construire et faire évoluer les profils des utilisateurs.

2. Les étapes du cycle de vie de la requête où l'on intègre le profil utilisateur dans :

(a) ***la phase de reformulation de la requête*** afin de mieux cibler le contexte de la recherche de l'utilisateur,

(b) ***la phase de réduction de l'espace de recherche*** pour restreindre l'espace de recherche aux documents qui ciblent les besoins de l'utilisateur,

(c) ***la phase d'appariement*** pour calculer la pertinence des documents en fonction des caractéristiques spécifiques de l'utilisateur,

(d) ***la phase de présentation des résultats*** pour restituer les informations selon le contexte et les préférences de l'utilisateur.

II.4. Modélisation du profil utilisateur

L'introduction de la dimension utilisateur dans un processus d'accès à l'information, mérite, voire nécessite une réflexion sur la modélisation de l'entité **utilisateur**.

La fiabilité ou qualité des profils est en effet d'une importance bien connue dans le domaine de la modélisation utilisateur (*User modeling*) **[A. Kobsa 2001].** En effet, on constate que l'une des principales raisons du manque de performances des techniques de personnalisation est typiquement l'application d'un profil utilisateur hors contexte **[S. Gauch & al. 2007].**

Les utilisateurs peuvent avoir des préférences générales, récurrentes et stables. Cependant, l'ensemble des informations contenues dans le profil ne sont pas forcément appropriées à toutes les situations de recherche.

Le plus souvent, les systèmes n'utilisent seulement qu'un sous-ensemble de ces informations, qu'ils supposent pertinents pour la recherche en cours.

Dés lors, le choix du profil adéquat constitue la principale réflexion lors de la mise en œuvre du SRIP.

De ce fait, les questions fondamentales posées pour modéliser le profil utilisateur sont le «*Quoi* », le « *Comment* » et le « *Quand* » **[G. Amato & al. 1999]:**

Quoi ?

Quelles propriétés informationnelles caractérisent l'utilisateur ?

Quelle structure informationnelle utiliser pour représenter l'utilisateur ?

Comment ?

Comment collecter les informations du profil ?

Comment construire le profil de l'utilisateur ?

Comment détecter le contexte, le but de la recherche et les besoins à court/long terme de l'utilisateur ?

Comment adapter le profil à l'évolution de l'utilisateur lui même ?

Comment assurer la sécurité et la confidentialité des informations du profil ?

Quand ?

Quand faut-il faire évoluer le profil de l'utilisateur ?

Plusieurs recherches **[kostadinov, 04] [Kien, 06]** distinguent principalement huit dimensions capables d'accueillir la plupart des informations caractérisant un profil. Ces dimensions sont les données personnelles, les centres d'intérêts, l'ontologie du domaine, la qualité attendue des résultats délivrés, la customisation, la sécurité et la confidentialité, le retour de préférences et les informations diverses.

1. Données personnelles : elles constituent la partie statique du profil et sont constituées des données démographiques, professionnelles, économiques et sociales d'un utilisateur. Ces données décrivent principalement l'identité de la personne, et son domaine d'activité. Elles sont relativement stables et ne jouent souvent aucun rôle dans le processus de recherche d'information.

2. Centres d'intérêts : le centre d'intérêt exprime le domaine d'expertise de l'utilisateur ou son périmètre d'exploration. Il peut être défini par un ensemble de mots clés (concepts) ou un ensemble d'expressions logiques (requêtes). Dans de nombreuses approches l'importance de chaque concept est définie par une pondération des mots clés du centre d'intérêt. L'ontologie du domaine complète la définition du centre d'intérêt en explicitant la sémantique de certains termes. Par exemple, on peut explicitement définir que 'BD' signifie 'bases de données' dans le profil et non 'bande dessinée', que dans le contexte du profil 'client' et 'consommateur' sont synonymes.

Le centre d'intérêt peut être vu comme une présélection virtuelle qui réduit la masse d'informations à prendre en compte. On peut rapprocher le centre d'intérêt de la notion de vue en BD.

Par conséquent toute requête émise par l'utilisateur sera enrichi avec les mots clés ou les prédicats des requêtes définissant le centre d'intérêt. Le centre d'intérêt peut être corrélé avec

les données personnelles et s'enrichir par déduction de certaines informations comme nous l'avons vu précédemment.

3. Ontologie du domaine : Elle explicite la sémantique de certains termes employés par l'utilisateur dans son profil et donne par conséquence une meilleure interprétation et signification de ses centres d'intérêts selon le domaine et le contexte dans lequel il travaille.

Cette ontologie peut être spécifique à l'utilisateur et explicitement définie par lui ou générique relative à un domaine particulier et dont la terminologie est clairement définie dans un thésaurus par exemple ;

4. Qualité attendue des résultats délivrés : est un des facteurs clés de la personnalisation, elle permet d'exprimer des préférences extrinsèques sur 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; elle sera confrontée à la qualité effective produite par le système de recherche d'informations. Il faut noter que la qualité d'un produit informationnel ne se mesure pas toujours sur le produit lui-même, mais quelquefois sur sa source de production ou son processus de production.

5. Customisation : elle concerne l'adaptation et la personnalisation de l'interface selon les préférences et les commodités de l'utilisateur tel que les modalités de présentation des résultats et les choix esthétiques ou visuels de l'utilisateur, la quantité de résultats qu'il souhaite recevoir, etc.

6. Sécurité et confidentialité : La sécurité est une dimension fondamentale du profil. Elle peut concerner les données que l'on interroge ou modifie, les informations que l'on calcule, les requêtes utilisateurs elles-mêmes ou les autres dimensions du profil. La sécurité des données peut être exprimée par des niveaux de sécurité prédéfinis qui dépendent de la hiérarchie des vues autorisées, par le support d'identification ou de stockage (carte à puce, certificat web etc.) et par des moyens de transmission utilisés (protocoles, cryptage). Les niveaux de sécurité peuvent concerner différents types d'«objets» comme les catégories du profil, les résultats des requêtes, mais aussi un processus de traitement ou une fonction de

calcul. La sécurité du processus exprime la volonté de l'utilisateur de cacher un traitement qu'il effectue. Ceci peut être fait en définissant le degré de visibilité de certaines opérations.

7. Retour de préférences : il désigne le « feedback » utilisateur qui peut être explicite et expressément fourni par lui ou implicite à travers l'analyse de certaines informations récupérées ou dérivées à son insu ;

8. Informations diverses : il peut être parfois souhaitable de fournir certaines informations spécifiques selon l'exigence de l'application ou du contexte de travail.

Le schéma d'un tel découpage du profil est le suivant :

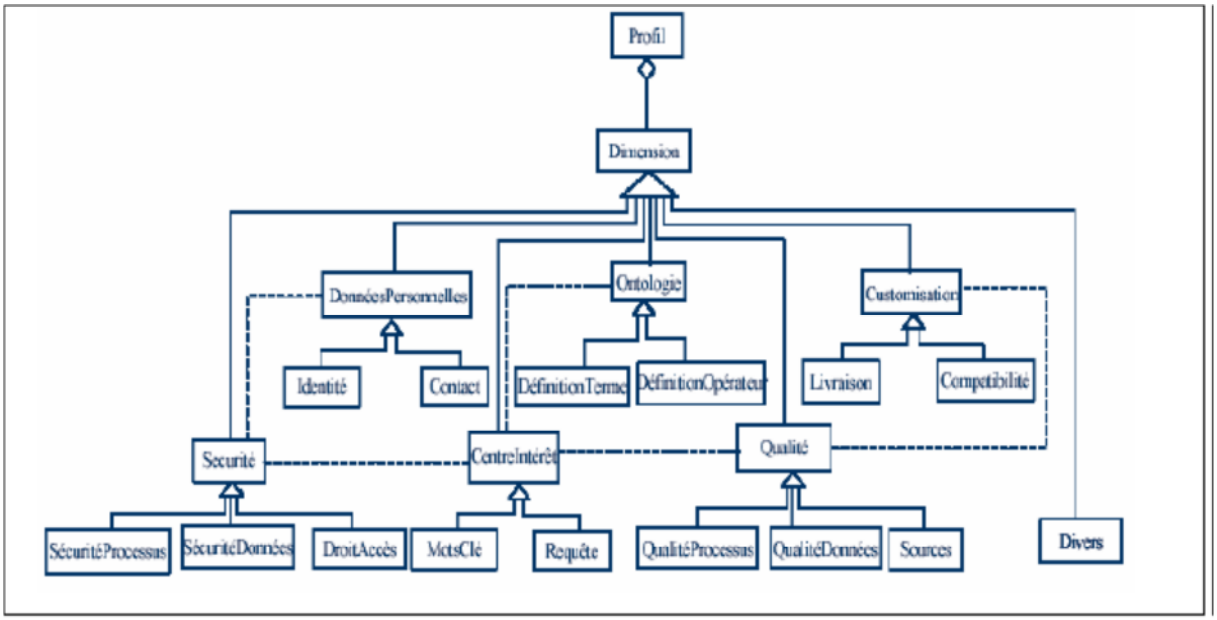


Figure II.2 : Dimensions et sous dimensions du modèle de profil

II.4.1. Représentation du Profil Utilisateur

Le profil de l'utilisateur n'a pas forcement de structure explicite qui le représente. Il peut être constitué de paquets divers d'informations qui traduisent une connaissance éparse sur l'utilisateur. Dans ce sens, la représentation des profils rejoint en grande partie la représentation de l'information dans le contexte de la recherche d'information. Il n'existe pas un modèle spécifique dédié à la représentation du profil de l'utilisateur.

Les modèles proposés puisent largement de ceux proposés en recherche d'information on en cite principalement cinq types de représentation : vectorielle, sémantique, connexionniste, multidimensionnelle et hiérarchique.

II.4.1.1. Représentation vectorielle

Ce type de représentation s'appuie généralement sur le modèle vectoriel proposé par Salton en 1971**[Salton, 71].** Le profil est représenté par un ou plusieurs vecteurs défini(s) dans un espace de termes obtenus implicitement ou explicitement à partir de plusieurs sources d'information. Les coordonnées des vecteurs correspondent aux poids des termes dans le profil.

L'utilisation de plusieurs vecteurs permet la prise en compte de la diversité des centres d'intérêts et de leur évolution dans le temps. Ce type de représentation offre l'avantage de la simplicité de mise en œuvre.

Cependant cette représentation manque de structuration et pose le problème du classement par ordre d'importances des préférences et des centres d'intérêts de chaque utilisateur et ne mettent pas en évidence la dimension liées au temps marquant l'évolution des profils.

II.4.1.2. Représentation sémantique

Cette représentation est essentiellement basée sur l'utilisation d'ontologies qui sont utilisées pour mettre en relief les relations sémantiques qui relient les unités d'informations qui composent le profil utilisateur en apportant des solutions aux problèmes de dissémination et de synonymie. La gestion proposée dans ce contexte, est la construction hiérarchique de concepts plutôt qu'une liste de structures indépendantes, à partir d'information issues des fichiers *logs*.

II.4.1.3. Représentation connexionniste

C'est un type de représentation basé sur l'interconnexion de nœuds représentant les termes qui composent le profil **[Jenningd, 93].** Il offre le double avantage de la structuration et de la représentation associative permettant la prise en considération de tous les aspects représentatifs du profil utilisateur.

II.4.1.4. Représentation multidimensionnelle

C'est un type de représentation qui se veut global dans le sens ou il permet de capturer puis catégoriser l'ensemble des informations caractérisant le profil de l'utilisateur. Le profil utilisateur peut contenir plusieurs types d'information telles que les données démographiques, les centres d'intérêts, les informations historiques et d'autres.

Chaque type d'information est une dimension dans le modèle multidimensionnel **[Kien, 06].** Le profil est donc structuré selon un ensemble de dimensions, représentées selon divers formalismes. Cette représentation a l'avantage d'apporter une meilleur interprétation de la sémantique du profil utilisateur et jouit d'une totale indépendance de l'application qui l'utilise. Cependant, elle souffre d'une certaine ambigüité quand il faut interpréter les rôles de chaque dimension du profil dans le processus de personnalisation.

II.4.1.5. Représentation hiérarchique

Cette représentation préconise la représentation du profil utilisateur à travers la construction d'une hiérarchique de concepts au lieu d'un ensemble de domaines indépendants. Chaque catégorie de la hiérarchie représente la connaissance d'un domaine d'intérêt de l'utilisateur. De plus, la relation généralisation /spécification entre les éléments de la hiérarchie traduit d'une manière plus réaliste les centres d'intérêts de l'utilisateur et qui ne sont pas toujours indépendants les uns des autres.

SRIP (Système de Recherche d'Information Personnalisé) est un exemple de système qui utilise cette représentation. Il se base sur la sélection dans une ontologie générale de nœuds estimés correspondant aux intérêts de l'utilisateur pour représenter le profil **[Tamine, 05].**

II.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.

II.4.2.1. Acquisition des données utilisateurs

Cette phase consiste à collecter les informations pertinentes pour instancier le profil de l'utilisateur.

Le processus d'acquisition des données de l'utilisateur implique différentes formes de diagnostic ou d'évaluation. Ce processus peut collecter ces informations soit directement à partir la machine de l'utilisateur (côté client) ou à partir de l'application (côté serveur).

Ce processus d'acquisition peut être explicite et/ou implicite :

II.4.2.1.1. L'acquisition explicite

L'acquisition explicite des données utilisateur repose principalement sur les techniques de feedback explicite largement utilisées dans les systèmes de filtrage et de reformulation de requêtes par réinjection de pertinence. Ces techniques sont utilisées dans le cadre de la RI personnalisée dans le but de maintenir et faire évoluer le profil de l'utilisateur. L'acquisition explicite permet à l'utilisateur de saisir manuellement ses domaines d'intérêts ou alors juger de façon explicite les documents renvoyés pour une requête servant à la construction de son profil **[Daoud, 09]**.

II.4.2.1.2. L'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 [**Zemirli, 08**].

II.4.2.1.3. Discussion : acquisition explicite vs acquisition implicite

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

En 2000, Quiroga et Mostafa **[L. Quiroga 2000]** ont comparé les performances du feedback explicite et implicite en analysant les résultats d'un système de filtrage d'information avec 18 utilisateurs sur une collection de test de 6.000 Mo disques de santé médicale classifiés en 15 domaines différents.

Chaque utilisateur a utilisé le système durant 15 sessions de recherche.

Ils ont obtenu une amélioration de la précision d'une valeur approximative à 68% lors de l'utilisation de profils utilisateurs construits à partir de la combinaison des deux approches de feedback.En outre, les résultats basés uniquement sur le feedback explicite ont produit une

précision maximale autour de 63% et de 58% comparativement à ceux basés uniquement sur l'implicite.

Ces différences s'avèrent statistiquement significatives. Ces expérimentations suggèrent que l'utilisation d'un profil explicitement construit ou un profil établi par la combinaison des deux approches produit de meilleurs résultats qu'une recherche basée sur un profil uniquement construit de manière implicite.

Cependant, contrairement aux précédents résultats, **[R. W. White 2001]** n'ont pas trouvé des différences significatives entre les profils construits implicitement et explicitement.

En effet, ils ont comparé deux systèmes de recherche sur le *web*, chacun basé sur le feedback explicite et implicite respectivement. Les expérimentations effectuées ont été menées par 16 utilisateurs ayant pour objectif de retrouver l'information sur le *web* en réponse à des requêtes spécifiques portant sur quatre domaines d'intérêts. L'accomplissement réussi de la recherche, la durée, ainsi que le nombre de pages résultant consultés pour chaque recherche ont été choisis comme métriques d'évaluation des performances des systèmes. Les utilisateurs avec des profils implicitement construits ont consulté approximativement 3;3 pages retrouvées pour chaque recherche, ce qui est d'avantage plus élevé que les 2;5 pages consultées dans le cas des profils explicites.

Les auteurs ont conclu que les approches explicites et implicites étaient similaires car ces différences n'étaient statistiquement pas significatives.

Plus récemment, **[J. Teevan & al.2005]** ont évalué diverses sources d'informations pour la construction du profil utilisateur tels que les pages *web* visitées, les *messages* échangés et l'ensemble des documents stockés sur la machine de l'utilisateur. Ils ont testés plusieurs profils utilisateurs construits à partir de différentes collections de tests issues par exemple uniquement des documents récemment enregistrés, des pages *web* uniquement, et combinaison des différentes sources.

En outre, ils ont construit deux profils utilisateur à partir de l'historique de recherche de l'ensemble des requêtes (préalablement soumises) et à partir de l'ensemble des domaines d'intérêts consultés lors de la navigation de l'utilisateur. Ils ont constaté que les performances

de recherche augmentent corrélativement avec la quantité d'information utilisée pour construire le profil utilisateur.

Ces différentes études, prises dans leur ensemble, suggèrent que les méthodes implicites sont sensiblement plus performantes que l'approche explicite pour la construction des profils utilisateur.
Le principal avantage de l'approche implicite est qu'elle est totalement transparente dans le processus d'acquisition des données pour l'utilisateur.

De plus, le profil peut être mis à jour plus fréquemment que lors d'une construction explicite. Ces mises à jour permettent la coordination de l'évolution du profil de l'utilisateur à long terme.

Néanmoins, l'inconvénient inhérent aux techniques implicites est qu'elles sont incertaines du fait de l'incertitude sous jacente aux comportements des utilisateurs et que les prédictions issues de l'observation de ces comportements sont difficilement quantifiables.

II.4.2.2. 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*.

II.4.2.2.1. Extraction d'ensemble de termes

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

systèmes appliquant cette approche de construction, on peut citer les systèmes ***WebMate*** **[Chen, 98]** et ***Alipes*** **[Widyantoro, 99]**.

II.4.2.2.2. Extraction de réseau de termes

Les termes sont extraits des documents jugés par l'utilisateur. Néanmoins, à la différence des approches précédentes, où les termes forment des vecteurs, les techniques de construction sémantique intègrent ces termes dans un réseau de nœuds. La construction des profils nécessite l'exploitation de relations préexistantes entre les termes et les concepts, tels que WordNet dans le cas du système SiteIF **[Stefani, 98],** ou manuellement construites tel que celui effectué par *WIFS* **[Micarelli, 04].**

Dans les approches élémentaires, chaque utilisateur est représenté par un seul réseau sémantique dans lequel chaque nœud contient un mot-clé unique. Lorsqu'un terme est présent dans le réseau, le poids de son nœud est augmenté ou diminué selon le feedback de l'utilisateur. Si le terme n'apparaît pas dans le réseau, un nouveau nœud est créé.

Les poids dans le réseau sont périodiquement réévalués à chaque mise à jour dans le but de modéliser les changements des centres d'intérêts de l'utilisateur à long terme. En outre, les concepts qui ne sont plus d'actualité peuvent être supprimés du réseau.

II.4.2.2.3. Extraction de concepts

L'approche de construction s'effectue de manière générale comme suit:

1. Identifier les concepts et niveaux de l'ontologie à exploiter : l'objectif étant d'extrait un sous ensemble de concepts représentant le profil général. Dans la plus part des travaux la ressource sémantique n'est pas exploité dans sa totalité. Certes, l'utilisation de tous les concepts de la hiérarchie permet d'obtenir des profils utilisateurs assez précis, pouvant couvrir un grand nombre de centres d'intérêts. Cependant, la difficulté de cette approche, se situe à juste titre au niveau de la profondeur de la hiérarchie d'ODP4 et la richesse des concepts. De ce fait, en général les systèmes extraient un nombre réduit de concepts à partir des premiers niveaux de la hiérarchie.

2. Extraire les centres d'intérêts de l'utilisateur par analogie aux concepts de l'ontologie : cette phase correspond à la phase de construction proprement dite. En ce sens où le profil de

chaque utilisateur est instancié à partir du profil général (la ressource sémantique) sur la base des informations collectées de l'utilisateur.

De manière générale, l'approche consiste en premier à associer à chaque catégorie sémantique de l'ontologie un ensemble de documents représentatifs du concept, puis à projeter sur le profil général, les documents (ou descripteur de documents) issue des différents feedback utilisateurs pour extraire les concepts représentant le profil de l'utilisateur. Parmi les différentes approches pour la construction de ces profils conceptuels, l'approche de coloration d'arbre qui est utilisée dans le système Persona **[Tanudjaja, 02]**.

II.4.3. Exploitation du profil utilisateur

Le domaine d'application conditionne fortement les techniques de personnalisation employées par le système.

Cela a en effet un impact direct sur l'exploitation du profil dans la chaîne d'accès à l'information et par conséquent sur les mécanismes mis en œuvre.

Les principales interrogations posées concerne le « Quoi », le « Comment » de la mise en œuvre :

Quoi ?

Quels services de personnalisation proposée : de la recommandation et/ou du filtrage, de l'aide à la navigation, un assistant personnel de recherche ?

Dans quelles étapes du cycle de vie de la requête faut-il intégrer le profil ?

Quelles informations du profil exploiter lors de l'accès à l'information ?

Comment ?

Comment intégrer le profil de l'utilisateur dans le processus de personnalisation ?

Comment évaluer l'impact de la personnalisation sur le processus de recherche ?

II.4.3.1. Les modèles d'accès personnalisé à l'information

Afin de mieux cibler les besoins et les préférences des utilisateurs, il faut adopter une manière efficace qui peut servir comme outil afin de répondre aux attentes de l'utilisateur. Les modèles de personnalisation utilisés pour servir l'utilisateur en lui retournant les résultats les plus pertinents. Dans ce cadre, **[KOS, 04]** a proposé quatre modèles de personnalisation qui sont couramment utilisés aujourd'hui à savoir la reformulation de la requête, appariement personnalisé de l'information requête-document, le ré-ordonnancement des résultats et la recommandation.

II.4.3.1.1. Modèle de la reformulation de requêtes

Selon KOS**,** ce service consiste à enrichir la requête par un ensemble de prédicats contenus dans son profil pour mieux cibler les informations dont il a réellement besoin. D'un coté, le fait d'ajouter des prédicats aux requêtes permet une meilleure exploitation des capacités des serveurs qui ont souvent des moyens d'exécution optimisés, mais d'un autre coté, c'est un processus très délicat à cause du risque d'insertion d'informations incorrectes qui provoqueraient le retour de résultats sans aucun intérêt**.**

II.4.3.1.2. Modèle d'appariement personnalisé de l'information requête-document

Le filtrage d'information est un processus permettant à partir d'un large volume d'informations dynamiques, d'extraire et de présenter les seuls documents intéressants ayant des centres d'intérêts relativement semblables appelés profils **[Boughanem, 01].**

Le principe de base de ce service est d'exécuter la requête sans prendre en compte la personnalisation puis ensuite appliquer un post traitement sur le résultat. Le filtrage peut être fait soit en appliquant des requêtes supplémentaires sur le résultat retourné, soit en traitant chaque élément séparément afin d'étudier sa pertinence.

L'avantage du filtrage des résultats est sa simplicité parce qu'il ne nécessite aucune modification du fonctionnement des fournisseurs d'information. Tout le traitement est fait après l'exécution de la requête. Les inconvénients sont le volume de données échangées entre le serveur et le client et le risque d'élimination d'éléments pertinents. Le filtrage des résultats se fait le plus souvent sur le côté client donc on a un gros transfert de données et en plus il y a

un risque de suppression d'objets pertinents par le fait que les éléments du résultat qui sont jugés non intéressants sont éliminés définitivement.

II.4.3.1.3. Modèle de ré ordonnancement des résultats de recherche

Un système de recherche d'information (SRI) retourne à l'utilisateur une liste de documents ordonnés selon leur degré de pertinence en réponse à sa requête. Cette requête constitue la principale source d'évidence pour déterminer la pertinence des documents. Le problème posé est que l'utilisateur n'ayant aucune (voire peu de) connaissance sur la collection de documents et de l'environnement de recherche, il lui est difficile de formuler des requêtes claires et appropriées, permettant de cibler la recherche uniquement aux documents pertinents **[Ruthven, 03].** Par conséquent, les documents pertinents se trouvent mélangés, lors de la présentation des résultats, aux documents non pertinents.

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

II.4.3.1.4. Modèle de recommandation d'objet du résultat

La technique de recommandation est un processus qui consiste à proposer à l'utilisateur des éléments correspondant à ses préférences ou en se servant de l'expérience des autres utilisateurs. Deux grandes approches sont proposées dans la littérature :

- L'approche basée sur la notion de profil utilisateur : elle consiste à proposer aux utilisateurs ayant un profil similaire. Parmi les techniques les plus utilisées, nous citons le filtrage collaboratif ou les techniques d'apprentissage automatique;

- L'approche basée sur la technique de fouilles des données : elle consiste à proposer des recommandations en s'appuyant sur l'observation et l'analyse du comportement d'un utilisateur ou d'autres utilisateurs lors de la navigation sur le web.

La recommandation d'éléments a l'avantage de pouvoir répondre aux demandes des nouveaux utilisateurs et de fournir aux utilisateurs des résultats sans les borner dans leur choix.

II.5. Évolution du profil utilisateur

La gestion de l'évolution du profil utilisateur est un processus complémentaire à la construction d'un profil utilisateur qui consiste 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.

L'évolution du profil utilisateur se fait souvent selon un processus incrémental basé sur l'addition de nouvelles informations dans 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.

Le premier représente les centres d'intérêts liés aux activités de recherche courantes de l'utilisateur.

Le second représente les centres d'intérêts persistants de l'utilisateur et issus de son historique de recherche tout entier.

II.5.1. 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.

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

Par conséquent, ce profil permet d'adapter efficacement le processus de RI aux besoins en information spécifiques de l'utilisateur.

II.5.2. Évolution du profil utilisateur à long terme

Le profil utilisateur à long terme modélise des centres d'intérêts généraux, persistants, ou récurrents.

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.

Plusieurs systèmes développés en RI personnalisée modélisent un profil utilisateur à long terme propre à chaque utilisateur. Parmi ces systèmes, les moteurs de recherche sur Internet *Google's Alerts*, et *Google's personalized search 1.1* et *Yahoo My Web*.

Dans les systèmes ou le profil utilisateur est spécifié selon un mode d'acquisition explicite des métadonnées par l'utilisateur, l'évolution de ce profil consiste à ajouter explicitement des domaines d'intérêt ou de les supprimer. *Google's Alerts* et la première version de *Google's personalized search 1.1* gèrent des profils utilisateurs génériques incluant les domaines de recherche stables gérés par les utilisateurs.

II.5.3. Approches de délimitation des sessions de recherche

L'évolution du profil utilisateur à court terme nécessite des techniques d'identification et de collecte des informations utiles et fortement liées aux activités de recherche courantes de l'utilisateur.

Ces techniques se basent souvent sur des mécanismes de délimitation des sessions de recherche définies par un intervalle de temps ou une séquence de requêtes liées à un même besoin en information.

D'après **[X. Huang & al.2004],** une session de recherche est définie par un groupe de requêtes soumises par un même utilisateur pour une même tâche de recherche.

Sur un intervalle de temps, un utilisateur peut faire une ou plusieurs sessions de recherche. Dans le but de clustériser les sessions de recherche, plusieurs approches ont ´été introduites dans la littérature.

Ces approches peuvent être classifiés en trois catégories : les approches basé temps, les approches basé contenu et les approches sémantiques.

II.5.3.1. Les approches basé temps

Les premières approches de clustérisassions des sessions de recherche sont basées sur la spécification d'un intervalle de temps moyen pour une session, appelé *Timeout* ***[B.hugues et all 2006]*** .

Dans ce type d'approches, la session est définie par une séquence de requêtes telle que l'intervalle de temps séparant deux requêtes successives ne dépasse pas un certain seuil. L'analyse est faite sur deux fichiers logs et montre qu'un intervalle de temps entre 10 et 15 minutes est identifiés comme le seuil optimal d'identification des sessions de recherche basé temps.

Cette méthode souffre du problème de la spécification du meilleur intervalle de temps pour identifier une session.

En effet, des utilisateurs différents peuvent avoir des comportements de navigation différents et l'intervalle de temps représentant le seuil d'identification des sessions peut être significativement différent.

De même, cet intervalle peut varier entre les sessions de recherche pour un même utilisateur. D'autres approches dédiées à l'analyse des fichiers log des moteurs de recherche identifient les sessions par regroupement des données des utilisateurs sur la base de l'adresse IP, les cookies et un intervalle de temps optimal **[I. Kang & al. 2004].**

Une méthode d'identification transactionnelle appelée « *reference length* », est proposé dans **[F. Crestani & al.2007].**

Cette méthode assume que le temps de lecture d'une page est corrélé au fait que la page est une page de contenu qui intéresse l'utilisateur ou une page auxiliaire.

Une nouvelle session est détectée à chaque détection d'une page « contenu ».

La limitation de cette méthode réside par le fait qu'un utilisateur peut s'intéresser à plus d'une page pour un même but de recherche.

Une autre méthode d'identification des sessions, appelée *maximal forward reference*, a été proposée dans **[M.-S. Chen & al.1998].**

Dans cette méthode, une session est définie sur un intervalle de temps par un ensemble de pages agrégées à partir de la première page visitée par l'utilisateur pour une séquence de requêtes jusqu'à ce qu'une page soit revisitée dans la session.

La limite de cette méthode est qu'elle traite une session par un ensemble de pages durant un intervalle de temps sans considérer la séquence des *clics* sur les pages visitées.

L'approche de clustérisassions des sessions proposée dans **[G. I. Webb & al.2001]** se base sur le principe de l'alignement séquentiel et prend en compte l'ordre des pages visitées dans une session dans le calcul de similarité des sessions.

II.5.3.2. Les approches basé contenu

Ces approches sont basées sur des mesures de similarité textuelle qui se catégorisent en des mesures basé mots clés ou phrases ou alors des mesures basées sur la distance d'édition des chaines de caractères entre deux requêtes successives **[R. W. White & al. 2003].**

Les mesures basées mots clés consistent à calculer le nombre de termes présents en commun entre deux requêtes successives p et q.

Cette similarité est définie par la formule suivante :

$similarity_{keyword}$(p, q) = KN (p, q)/Max (kn(p), kn(q)) (2.4)

Où

kn : est le nombre de termes présents dans une requête,

kn (p, q) : est le nombre de termes présents simultanément dans les deux requêtes p et q.

Des dérivations de cette formule consistent à calculer une similarité où les termes de la requête sont pondérés et peut être étendue pour calculer une similarité plus précise entre les requêtes basée sur les phrases plutôt que des termes uniques.

Dans cette approche étendue, l'unité élémentaire représentant partiellement la requête n'est plus un terme mais un ensemble de termes groupés selon des règles syntaxiques.

Ceci augmente la similarité entre deux requêtes ayant une phrase en commun en réduisant le nombre des unités élémentaires (phrase) différentes.

D'autres mesures consistent `a calculer la distance d'édition des chaines de caractères entre deux requêtes successives **[D. Gusfield 1997].**

Cette mesure est inversement proportionnelle au nombre d'éditions nécessaires (insertion, suppression, etc.) à unifier deux chaines de caractères (requêtes).

La similarité entre deux requêtes p et q, est calculée selon la formule suivante :

$similarity_{edit}(p, q)$ = 1 – EditDistance (p, q) (2.5)

II.5.3.3. Les approches sémantiques

Ces approches sont basées sur des mesures de similarité sémantiques qui se catégorisent en des mesures basées sur le *feedback* utilisateur **[R. W. White & al.2003]** et des mesures basées sur l'information mutuelle **[L. Tamine 2008].**

Les mesures basées sur le *feedback* utilisateur **[R. W. White & al.2003]** consistent à calculer le nombre de pages visitées en commun pour deux requêtes successives.

L'intuition derrière cette mesure est que deux requêtes ayant des documents en commun visitées par l'utilisateur partagent le même sujet.

Cette mesure permet de grouper des requêtes sémantiquement liées dans une même session. La mesure de similarité entre deux requêtes p et q, est calculée selon la formule suivante :

$similarity_{feedback}$ (p, q) = RD (p, q) / Max (rd(p), rd(q)) (2.6)

Où

RD : est le nombre de documents *cliqués* communs entre les deux requêtes,

rd(p) : est le nombre de documents *cliqués* pour une requête p.

Une mesure de similarité plus élaborée dérivée de la mesure précédente est proposée dans **[R. W. White & al.2003]** et a pour but d'intégrer en plus du *feedback* utilisateur, une distance conceptuelle entre les documents cliqués communs entre deux requêtes.

La distance conceptuelle entre deux documents est calculée sur la base d'une hiérarchie de concepts (Encarta) dans laquelle chaque document de la collection est classifié dans le concept correspondant.

Le système proposé dans **[L. Tamine 2008]** intègre une mesure de similarité sémantique qui consiste à calculer le nombre de documents indexés par les termes provenant des deux requêtes successives.

Le but dans cette étude est de développer un SRI basé-session où le contexte est représenté par l'ensemble de requêtes et ses résultats associés dans une même session de recherche.

II.6. Évaluation d'un SRIP

L'évaluation des SRI est depuis le début des travaux sur la RI un des piliers de l'évolution de ce domaine.

La qualité de l'évaluation est d'une importance capitale puisqu'elle permet de discriminer les différents modèles.

Nous présentons dans cette section une synthèse des approches d'évaluation utilisées dans le cadre de l'accès personnalisé.

Nous décrivons en premier lieu, le protocole d'évaluation standard TREC (Text Retrieval Conference) dédié à la RI traditionnelle.

En second lieu, nous dressons un bilan des limites du protocole TREC 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 définirons ensuite par un aperçu de quelques travaux de référence.

II.6.1. 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 aux États-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.

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.

II.6.1.1. 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).

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.).

II.6.1.2. Collections de test

Les collections TREC sont de l'ordre de quelques giga-octets et de quelques centaines de giga-octets pour les VLC (Very Large Collections et TB Terabyte).

Les documents sont issus de différentes sources dont essentiellement la presse écrite tel que le Wall Street Journal mais également des documents *web*.

Ces données sont disponibles sur le serveur du NIST.

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 (dernièrement) des documents informatiques mis sur Internet.

Il semble que certains soient (faiblement) structurés : présence d'un titre, indication des paragraphes, les autres documents ont des structures hétérogènes annotées de métadonnées. Il existe quatre dimensions de variation :

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

Les quelques documents plus longs sont des brevets d'environ 3 000 mots.

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

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

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

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 :

II.6.2. Protocoles d'évaluation pour l'accès personnalisé

Des différents éléments abordés dans la section précédente, force est de constater qu'il n'y a actuellement aucune tâche de personnalisation dans TREC.

Aucune collection de test standard n'a été construite à notre connaissance pour évaluer l'efficacité de l'accès personnalisé à l'information.

De telles collections contiendraient divers éléments du contexte liés à l'utilisateur directement (historique de la recherche, centres d'intérêt, expertise etc.) ou à la session de recherche (but de la recherche, tâche, etc.).

<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>

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

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

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.

II.6.2.1. 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 **[L. Tamine & al.2008].**

1. la mesure RR (Relative Relevance).

La mesure RR **[P. Borlund & al.1998]** a pour objectif de considérer différents types de pertinence (pertinence non binaire) dans l'évaluation de l'efficacité d'un système d'accès contextuel à l'information. Cette mesure quantifie le degré de concordance entre les types de jugement de pertinence émis dans le cas de deux ensembles de jugements (soit R_1 et R_2) associés à une même liste de documents qui constitue les résultats d'une session de recherche.

En pratique, R_1 correspond généralement aux scores de pertinence algorithmique retournés par un SRI et R_2 à des scores de pertinence contextuelle correspondant à un type de pertinence donné : situationnelle si elle est exprimée par un utilisateur,

Thématique si elle est exprimée par un assesseur etc. La valeur de corrélation entre R_1 et

R_2 est généralement calculée en utilisant une mesure du cosinus ; elle quantifie globalement, la capacité du système à prédire le type de pertinence contextuelle considéré.

A la différence de la mesure classique de précision, cette mesure permet de considérer les différents types de pertinence ; néanmoins, elle pose un problème lors de l'évaluation comparative entre différents algorithmes de recherche voire entre différents SRI **[P. Borlund 2003].**

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

2. les mesure CG (Cumulative Gain) et DCG (Discount Cumulative Gain)

Les mesures CG et DCG 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 :

$$\mathrm{CG[i]}=\begin{cases} \mathrm{G[1\,]}, si\ i=1 \\ \mathrm{CG[i-1\,]}+G[i],\ sinon \end{cases} \qquad \textbf{(2.7)}$$

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

$$\mathrm{CG[i]}=\begin{cases} \mathrm{G[1\,]}, si\ i=1 \\ \mathrm{CG[i-1\,]}+G[i]/\log_{q},\ sinon \end{cases} \qquad \textbf{(2.8)}$$

Comparativement à la mesure CG, la mesure DCG permet d'atténuer le gain de pertinence apporté par un document en fonction du rang associé.

Ceci rejoint en effet l'hypothèse évidente que plus le rang d'un document est élevé, moins il est probable que l'utilisateur l'examine et donc moins il est à l'origine d'un gain effectif de pertinence.

3. La mesure GRP (Generalised Recall and Precision)

La mesure GRP **[K. Jarvelin & al.2000]** est également une mesure orientée position qui généralise les mesures classiques de rappel et précision en considérant une pertinence graduelle.

Le rappel généralisé (GR) et la précision généralisée (GP) sont calculés comme suit :

$$GP = \sum_{d \in R} r(d)/|R| \qquad \textbf{(2.9)}$$

$$\mathrm{GR} = \sum_{d \in R} r(d) / \sum_{d \in R} r(d) \qquad \textbf{(2.10)}$$

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.

II.6.2.2. 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 :

- **Étape 1 : Évaluer 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 **[J. Chaffee & al.2000; S. Dumais & al.2003].** 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 **[M. J. Pazzani & al.1996]** peuvent être utilisées où des mesures de convergence de l'algorithme **[F. Liu & al.2004].**

- **Étape 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 **[F. Liu & al.2004; S.Speretta & al.2004; S. Gauch & al.2003].**

Dans le cas de l'utilisation de mesures agrégées de rappel/précision, un référentiel est généralement construit sur la base de l'ensemble des documents pertinents jugés par l'ensemble des utilisateurs pour chaque requête.

L'utilisation de mesures orientées rang évite l'utilisation d'un tel référentiel.

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

1. Dans les travaux de **[F. Liu & al.2004]** la personnalisation consiste en la désambiguïsation de la requête de l'utilisateur en se basant sur le profil de l'utilisateur.

Pour valider leurs approches, ils effectuent les expérimentations se déroulant avec 7 utilisateurs. Chaque utilisateur soumet *n* requêtes à un *Google web Directory* en identifiant les catégories associées pour chacune des ces requêtes.

Pour établir la collection de test, chaque requête est exécutée selon 3 modes :

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

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

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

La collection de test référentiel est ensuite obtenue en regroupant pour chaque paire (utilisateur, requête) l'union de l'ensemble des documents jugés par l'utilisateur pour les 3 modes de soumission de la requête.

L'évaluation des profils appris passe par le test de l'efficacité des algorithmes utilisés pour la construction de ce profil en comparant les différents résultats obtenus par chacun des algorithmes.

Pour tester les performances de l'algorithme de construction du profil, ils augmentent à chaque apprentissage la taille des données (i.e., la taille de l'historique de recherche) en appliquant la stratégie de la *k*-fold Cross Validation, où *k* est positionné à 10.

Ils découpent ainsi la collection de test en 10 sous collections : 9 sous collections pour l'apprentissage et la 10*ieme* pour le test.

Puis, ils répètent l'expérimentation 10 fois.

A chaque *ième* expérimentation, *ième* sous collection est utilisée pour le test.

2. Les secondes expérimentations que l'on cite :

Le scénario général d'évaluation s'effectue avec 6 utilisateurs (étudiants de l'université du Kansas) durant 6 mois.

Chaque utilisateur soumet 45 requêtes à Google. Durant cette période, ils collectent pour chaque paire (utilisateur, requête) les 10 premiers résultats sélectionnés par l'utilisateur. Puis, ils découpent cette collection en deux sous-ensembles :

_ Du résultat des 40 requêtes, ils forment la collection pour l'apprentissage du profil.

_ Du résultat des 5 requêtes restantes, ils forment la collection de test.

Pour évaluer les profils appris, ils effectuent une série de tests en faisant varier la taille des collections d'apprentissage pour la construction du profil.

Les profils utilisateurs sont créés sur la base de 5, 10, 20, 30, puis 40 requêtes.

Ils construisent également des profils avec 30 requêtes et un nombre de 20 concepts issus d'ODP.

Par la suite, lors de l'évaluation de l'impact de l'intégration du profil dans le processus d'accès, la collection référentielle est obtenue par l'union des documents jugés pertinents par l'utilisateur pour l'ensemble des requêtes.

Ils mesurent l'exactitude des profils construits en effectuant des comparaisons statistiques entre le rang calculé par le système (sur la base de similarité avec le profil) et celui obtenu par Google pour chacun des 10 premiers résultats sélectionnés par l'utilisateur.

II.7. Conclusion

Dans ce chapitre nous avons présenté les principaux systèmes de personnalisation de recherches 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 différent approche et technique de modélisation du profil utilisateur, à savoir sa représentation, sa construction, son évolution au cours du temps ainsi son intégration dans les systèmes de recherches d'information.

Nous avons abordé les méthodes de détection automatique de sessions de recherches et les méthodes d'évaluation des systèmes d'accès personnalisées à l'information. Nous avons pu constater que pour faire assoir une personnalisation efficace, il lui faut une personnalisation efficace de l'utilisateur. Cette dernière dépend du model de représentation du profil de l'utilisateur, de sa construction et de son évolution au cours du temps .Ces élément sont a la base des différences de performances des systèmes de recherches d'information personnalisées.

Chapitre III : Evaluation expérimental délimitation des approches de session de recherche

III.1. Introduction

Ce chapitre est consacré à l'évaluation des différentes approches de délimitation des sessions de recherche en tenant compte de l'évolution du profil utilisateur. Notre démarche consiste à récupérer les interactions des utilisateurs à partir de leur historique de recherche. Cet historique est vu comme une source d'information, évoluant lors des différentes interactions de recherche de l'utilisateur, à partir desquelles on fait inférer les sessions de recherche selon les centres d'intérêts de l'utilisateur, le facteur temps ainsi qu'on tenons compte de la approche hybride (agrégation des deux approches précédentes) :

Dans la première partie du chapitre nous allons présenter notre démarche d'évaluation, ensuite les approches utilisées, enfin dans la dernière partie nous allons présenter les outils utilisés ainsi que les résultats finaux de toute la collection.

III.2. Notre démarche d'évaluation

III.2.1. L'acquisition des données utilisateur

Ces données correspondent aux documents jugés pertinents par l'utilisateur en observant son comportement face aux résultats de recherche (Lecture, Sauvegarde, Impression).

III.2.2. Processus de modélisation du profil utilisateur

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

(a) Pour chaque interaction nous récupérons les documents pertinents. A partir desquels nous définissons les centres d'intérêts,

(b) Inférence du centre d'intérêt :

Un centre d'intérêt est inféré lors d'une interaction de recherche. Plus précisément, il est construit à partir de l'ensemble D_r des documents pertinents, obtenus lors de cette dernière. Il est alors représenté par un vecteur de termes pondérés. Il est dénoté $c_k = \{(t_1,w_{1k}), (t_2,w_{2k}), ...,(t_i,w_{ik})\}$. Ces termes correspondent aux termes les plus pertinents des documents de l'ensemble D_r.

Le poids w_{ik} d'un terme t_i dans le centre d'intérêt c_k est calculé comme suit:

$$w_{ik} = \frac{1}{D_r} \sum_{dj \in D_r} w_{ij} \quad (3.1)$$

Où, w_{ij} représente le poids d'un terme t_i dans le document d_j. Il est obtenu par une variation normalisée de TF-IDF **[Robertson et al. 1995]** :

$$w_{ij} = tf_{ij} * \log \frac{N}{n_i} \quad (3.2)$$

Avec, tf_{ij} est la fréquence d'apparition du terme t_i dans le document d_j, N est le nombre total des documents et n_i est le nombre de documents qui contiennent le terme t_i.

III.2.3. Approche de délimitation des sessions de recherche

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

Une session de recherche est un ensemble d'interactions. Chaque interaction contient une requête ainsi que l'ensemble des documents jugés pertinents par l'utilisateur. Elle est caractérisée par une période [début d'interaction - fin d'interaction] .Notre objectif est de définir les sessions de recherche on se basons sur les trois approches basée : temps, contenu, temps et contenu.

L'idée de base est d'exploiter le profil de l'utilisateur pour déduire les sessions. En effet, le profil de l'utilisateur traduit son centre d'intérêt pour une ou plusieurs interactions. Notre travail consiste à avoir des collections de test de l'utilisateur réel afin de déduire le nombre de session de recherche pour chaque approche sachant que il y en a 3 approche

- Approche basée temps.
- Approche basée contenu.
- Approche hybride (basée temps et contenu)

Pour répondre à ce besoin nous avons opté pour une collection totale qui contient :

- 10 utilisateurs
- 81 documents

- 60 interactions (requête)

La figure 3.1 représente un extrait d'interaction de recherche de l'utilisateur N° 3

```
<?xml version="1.0" encoding="UTF-8"?>

<user num="3" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="user1.xsd">

<interaction duree="20 mn " num="1" starttime="10:15:00">

<query> " reseau de capteur " </query>

<results>

<result rank="1">

<url>http://irt.enseeiht.fr/beylot/enseignement/Capteurs.pdf</url>

<title> Réseaux de capteurs sans fil De la théorie à la pratique - ENSEEIH</title>

<snippet> Un "capteur" d'un réseau de capteurs. Capte toujours. Petit / Grand nombre. Limité en énergie.
Capable de calculer. Communicant. Architecture générale et. </snippet>

</result>

</results>

</interaction>
```

Figure 3.1 Interactions de l'utilisateur N° 3

III.2.3.1 Approche basée temps

Le principe de cette approche consiste à définir la durée moyenne d'une session de recherche et qui correspond aux moyennes des durées de toutes les interactions de tous les utilisateurs.

Durée d'une session de recherche= $\left[\sum_{n=1}^{10} \text{duree}_{\text{moyenne}}(\text{interaction n}) \right] / 10.$

Dans notre travail la durée moyenne vaut **25mn**.

L'extrait de fichier qui suit montre les interactions initiales faites par l'utilisateur N°9 avant application de la nouvelle durée.

<interraction num="1" duree_total="11 mn" >**écrire un chapitre avec lyx** </interraction>

<interraction num="2" duree_total="8 mn" > **modèle de system de recherche d'information personnalisé** </interraction>

<interraction num="3" duree_total="2 mn" > **acquittement larousse** </interraction>

<interraction num="4" duree_total="2 mn" > **arasem** </interraction>

<interraction num="5" duree_total="2 mn" > **cognitif** </interraction>

<interraction num="6" duree_total="13 mn" >**recherche de l'information cours - Recherche Google**

</interraction>

Figure 3.2 Interactions de l'utilisateur N° 9

Remarque : Après avoir passé par l'algorithme de l'approche basée temps on obtient les résultats ci-dessous.

<session num="16" duree="25 mn">

<interraction num="1" duree_total="11 mn" >**écrire un chapitre avec lyx** </interraction>

<interraction num="2" duree_total="8 mn" > **modèle de system de recherche d'information personnalisé** </interraction>

<interraction num="3" duree_total="2 mn" > **acquittement larousse** </interraction>

<interraction num="4" duree_total="2 mn" > **arasem** </interraction>

<interraction num="5" duree_total="2 mn" > **cognitif** </interraction>

</session>

<session num="17" duree="25 mn">

<interraction num="1" duree_total="13 mn" > **recherche de l'information cours – Recherche Google** </interraction>

<interraction num="2" duree_total="45 mn" >**recherche de l'information en 2019** </interraction>

</session> -→ Génération d'une nouvelle session pour compléter l'interaction num 2

Figure 3.3 Sessions utilisateur N° 9 selon l'approche basée Temps

III.2.3.1.1 Résultats de délimitation de sessions de recherche selon l'approche basée temps.

Nous présentons dans le tableau **III.1** un récapitulatif des résultats obtenu selon l'approche basée temps.

USER	NBR Interactions	Session basée temps
1	9	11
2	4	6
3	3	6
4	3	5
5	**3**	**3**
6	**3**	**3**
7	**4**	**4**
8	**4**	**4**
9	16	21
10	10	8

Tableau III.1: Résultats de l'approche basée temps.

III.2.3.1.2 Interprétation de l'approche basée temps

Les résultats obtenus montrent qu'un nombre élevé de sessions de recherche ont était effectuées par les utilisateurs N°(1,2,3,4,9) , ce qui n'est pas étonnant dans la mesure ou plusieurs interactions possèdent des durée dépassent la durée moyenne défini pour une session de recherche.

Par contre, le nombre de sessions de recherche est équivalents au nombre d'interactions réalisées par les utilisateurs N° (5, 6, 7,8) vu que la durée moyenne des interactions varie auteur des ***25 mn.***

Contrairement au nombre de sessions de recherche qui est inferieur au nombre d'interactions établi par l'utilisateur N° 10 puisqu'il il existe des interactions dont leur durées moyenne est inferieur a la durée moyenne de la session de recherche.

III.2.3.2. Approche basée contenu

Le Principe de cette approche consiste à détecter le changement de session de recherche lié au changement de centre d'intérêt.

Initialement, on indexe les documents pertinents d'une requête liée à une interaction de recherche. A partir des résultats obtenue on défini le profile qui correspond au vecteur de termes pondérés issu des documents pertinents. On calcule les scores de similarité entre le profile de deux (02) interactions qui ce succède selon la formule suivante :

$$\text{Cos}\left(\overrightarrow{vecteur\ p1}, \overrightarrow{vecteur\ p2}\right) = \left(\frac{p1.vecteur.dotproduct\ (p2.vector))}{\left|\overrightarrow{vecteur\ p1\ .getNorm}\right| Vecteur\ p2.getNorm}\right),$$

Si Score >= 0.6 similarité entre (p1, p2) ➔ Pas de changement de centre d'intérêt => les interactions appartiennent à la même session de recherche.

Sinon : il y a changement de centre d'intérêt => les interactions n'appartiennent pas à la même session de recherche.

L'extrait de fichier qui suit montre les interactions initial de l'utilisateur N° 9

<interaction num="1" duree_total="2 mn" >**cognitif** </interaction>

<interaction num="2" duree_total="2 mn" > **arrasem** </interaction>

<interaction num="3" duree_total="13 mn" > **recherche de l'information cours - Recherche Google**

</interaction>

<interaction num="4" duree_total="45 mn" > **recherche de l'information en 2019** </interaction>

<interaction num="5" duree_total="3 mn" > **lucen** </interaction>

<interaction num="6" duree_total="50 mn" > **model de recherche d'information personnalisée** </interaction>

Figure 3.4 Interactions utilisateur N° 9

Remarque : Après avoir passé par l'algorithme de l'approche basée contenu on obtient les résultats si dessous.

L'extrait de fichier qui suit montre les sessions obtenues par l'utilisateur N° 9 dans les interactions sont représentées comme suit :

<session num="1" duree_total="2 mn" >

<interaction num="1" duree_total="2 mn" >**cognitif** </interaction>

</session>

<session num="2" duree="2 mn" >

<interaction num="1" duree_total="2 mn" > **arrasem** </interaction>

</session>

<session num="3" duree_total="111 mn" >

<interaction num="1" duree_total="13 mn" > **recherche de l'information cours - Recherche Google**

</interaction>

<interaction num="2" duree_total="45 mn" > **recherche de l'information en 2019** </interaction>

<interaction num="3" duree_total="3 mn" > **lucen** </interaction>

<interaction num="4" duree_total="50 mn" > **model de recherche d'information personnalisée** </interaction>

</session>

Figure 3.5 Sessions utilisateur N° 9 selon l'approche basée Contenu

III.2.3.2.1 Résultats de délimitation de sessions de recherche selon l'approche basée contenu

Nous présentons dans le tableau **III.2** un récapitulatif des résultats obtenu selon l'approche basée contenu.

USER	NBR Interactions	session basée contenu
1	**9**	**7**
2	4	1
3	3	1
4	3	1
5	3	1
6	3	1
7	4	1
8	4	1
9	16	10
10	10	3

Tableau III.2 : tableau explicatif de l'approche basée contenu pour tous les utilisateurs

III.2.3.2.2 Interprétation des résultats précédents

Les résultats obtenus montrent qu'un nombre élevé de sessions de recherche a était effectuées par les utilisateurs N°(9,10), ce qui n'est pas étonnant dans la mesure où ils ont exécutés plusieurs interactions permettant ainsi d'observer le changement de centre d'intérêt.

Par contre, le nombre de sessions de recherche est presque équivalents au nombre d'interactions réalisées par l'utilisateur N° (10) vu que ce dernier a débuté sa recherche sans avoir une idée de recherche précise (une thématique de recherche).

Contrairement a l'unique session de recherche faite par les utilisateurs N° (2, 3, 4, 5, 6, 7, 8) vu qu'ils n'ont pas fait un nombre élevés d'interactions (vari autour de 2 à 3 interactions), ce qui ne permet pas d'observé le changement de session.

III.2.3.3. Approche hybride (basée temps et contenu)

Cette approche consiste à combiner les 2 approches précédentes. Autrement dit le changement de session de recherche et lié à la durée et au changement de centre d'intérêt utilisateur.

L'extrait du fichier qui suit montre les sessions obtenues par l'utilisateur N° 9 dont les interactions sont représentées comme suit :

→ <session num="26" duree="25 mn" >

<interaction num="1" duree_total="8 mn" > **recherche de l'information en 2019** </interaction>

<interaction num="2" duree_total="3 mn" > **lucen** </interaction>

<interaction num="3" duree_total="50 mn" > **model de recherche d'information personnalisée** </interaction>- → Génération d'une nouvelle session pour compléter l'interaction 3

</session>

→ <session num="27" duree="25 mn" >

<interaction num="1" duree_total="50 mn" > **model de recherche d'information personnalisée** </interaction> -→ Génération d'une nouvelle session pour compléter l'interaction 1

</session>

→ <session num="28" duree="25 mn" >

<interaction num="1" duree_total="50 mn" > **model de recherche d'information personnalisée** </interaction>

</session>

Figure 3.6 Sessions utilisateur N° 9 selon l'approche basée Temps et Contenu

III.2.3.3.1. Résultats de délimitation de session de recherche selon l'approche basée temps et contenu de tous les utilisateurs

Nous présentons dans le tableau **III.3** un récapitulatif des résultats obtenu selon l'approche basée temps et contenu.

USER	NBR Interactions	Session basée temps et contenu
1	9	13
2	4	6
3	3	6
4	3	5
5	3	3
6	3	3
7	4	4
8	4	4
9	16	29
10	**10**	**8**

Tableau III.3 : Résultats de l'approche basée temps et contenu

III.2.3.3.2. Interprétation de l'approche basée temps et contenu

Les résultats obtenus montrent qu'un nombre élevé de sessions de recherche ont était effectuées par les utilisateurs N° (1, 2, 3, 4,9), ce qui n'est pas étonnant dans la mesure où plusieurs interactions possèdent des durée dépassent la durée moyenne défini pour une session de recherche et le profil de certaine interactions (selon l'approche basée temps) n'appartiens pas à la même session (selon l'approche basée contenu) → **Génération de nouvelle session de recherche.**

Par contre, le nombre de sessions de recherche est équivalents au nombre d'interactions réalisées par les utilisateurs N° (5, 6, 7,8) vu que la durée moyenne des interactions varie auteur des ***25 mn*** ainsi que le centre d'intérêt n'est pas changé.

Contrairement au nombre de sessions de recherche qui est inferieur au nombre d'interactions établi par l'utilisateur N° 10 puisqu'il il existe des interactions dont leur durées moyenne est inferieur a la durée moyenne de la session de recherche et le centre d'intérêt n'est pas changé.

III.3. Outils de développement

Pour implémenter notre approche, nous avons été amenées à étendre certaines classes java de la plateforme lucene. Pour réaliser nos tests nous avons utilisés :

III.3.1. Eclipse IDE

Eclipse est un environnement de développement (IDE) historiquement destiné au langage Java, mais grâce à un système de plugin il peut être utilisé avec d'autres langages de programmation comme le PHP et le C/C++.

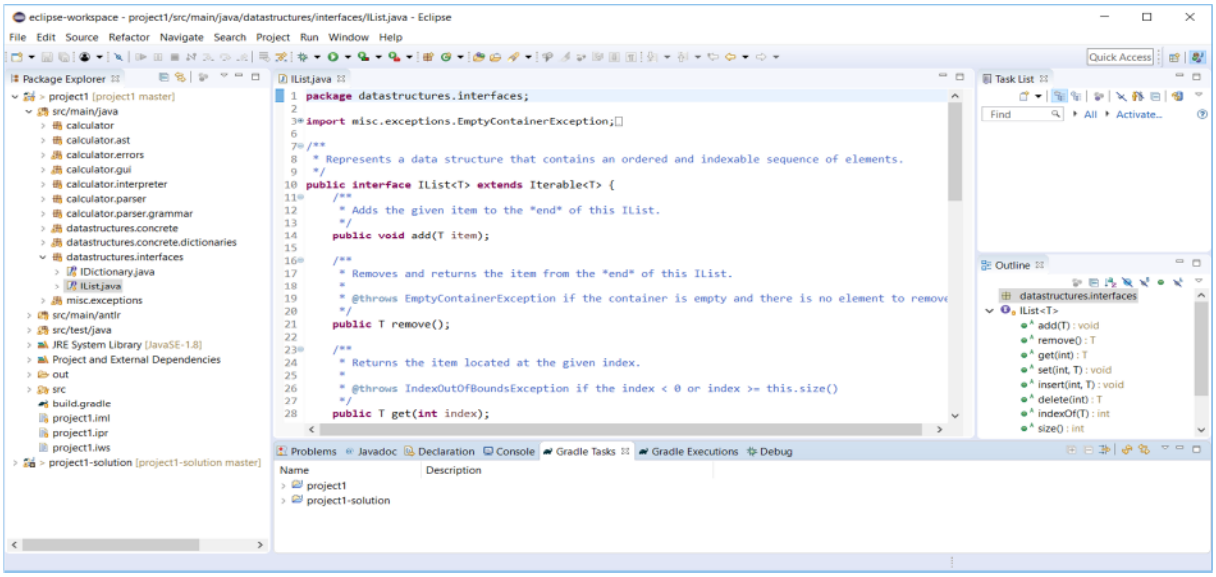


Figure III.7 : Interface de l'IDE éclipse

III.3.2. Langage java

Le langage java est un langage de programmation informatique orienté objet créé par James Gosling et Patrick Naughton, employés de Sun Microsystems. La particularité de java est qu'il est facilement portable sur d'autres systèmes d'exploitation tels que linux, Windows, Mac. . . avec peux ou aucune modification.

III.3.3. Editix

Editix est un éditeur XML qui fonctionne sur les plateformes Windows, GNU/Linux ou Mac OS X. En plus de la coloration syntaxique essentielle à l'écriture de documents XML, ce logiciel nous offre de nombreux outils qui nous seront utiles dans la suite de ce tutoriel comme par exemple la validation des documents.

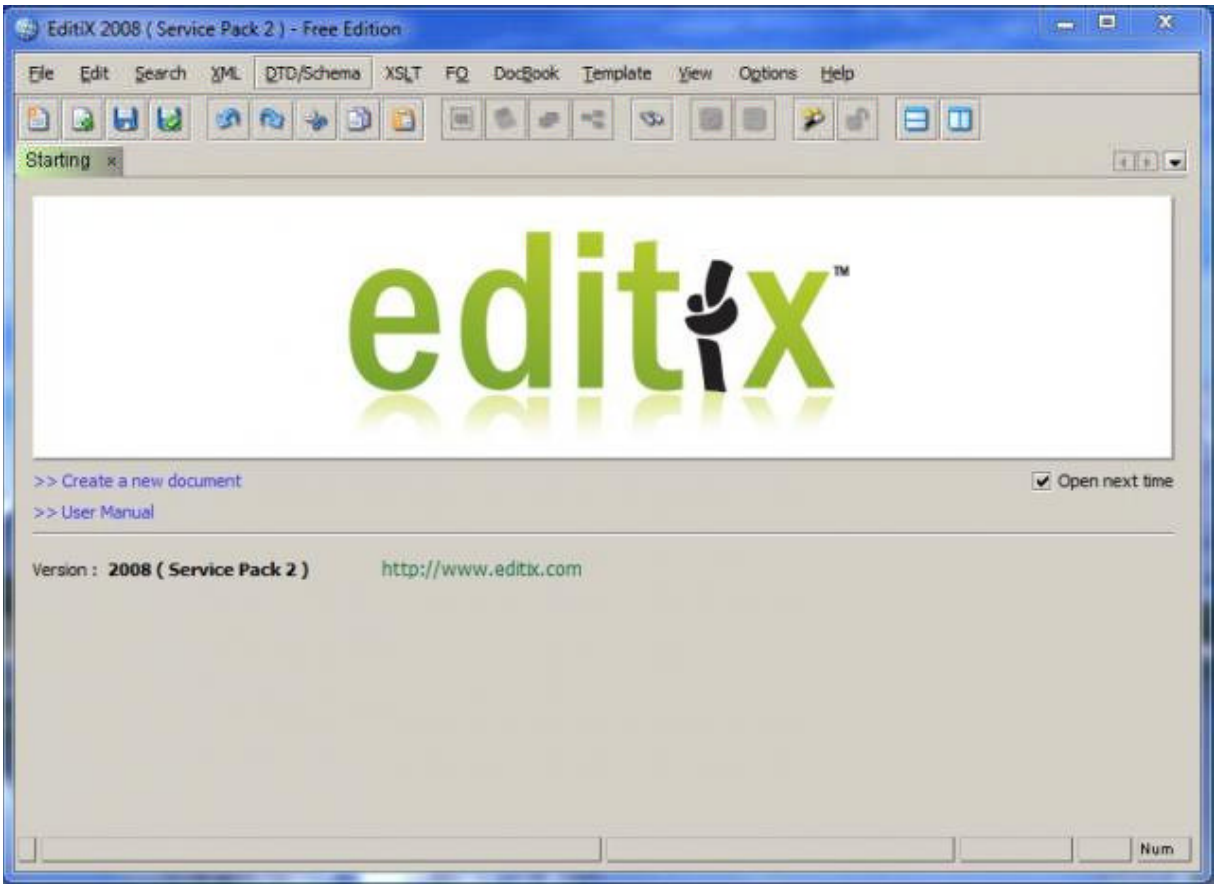


Figure III.8 : Interface de EDITIX

III.3.4. Lucene

Lucene est un moteur de recherche textuelle développé sous java par la fondation apache se concentrant sur l'indexation et la recherche de textes.

III.3.4.1. Architecture de Lucene

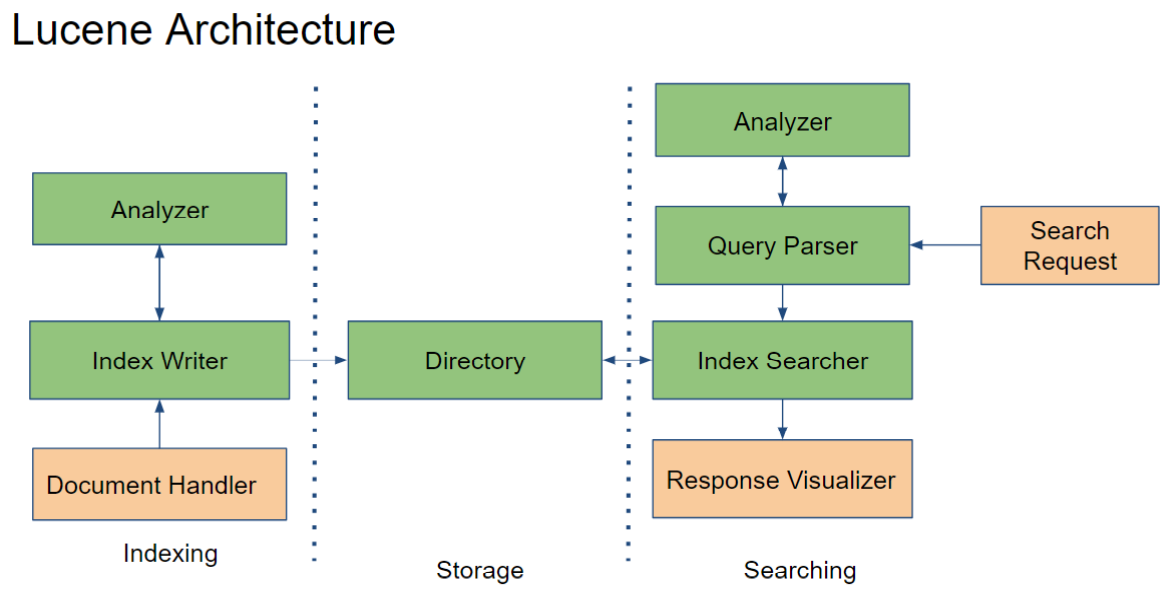


Figure III.9 : Architecture de Lucene

Lucene se découpe en 7 paquetages principaux :

— Org.apache.lucene.analysis : il contient du code afin de convertir du texte en élément indexable. Il contient la classe Analyzer qui permet d'extraire les mots importants pour l'index et supprimer le reste.

— Org.apache.lucene.document : contient des classes relatives aux documents, tel que la classe : Document. Cette classe représente un rassemblement de champs(Fields), ainsi les métadonnées sont indexées et stockées séparément comme des champs d'un document.

— Org.apache.lucene.index : il contient le code pour accéder aux index. On y trouve la classe IndexWriter qui permet de créer un index et d'ajouter des documents dans un index existant.

— Org.apache.lucene.queryparser : son rôle est d'analyser les requêtes afin dégénérer la requête sous forme d'objet query qui pourront ensuite être réutilisépar le parseur. On y trouve la classe la classe QueryParser qui est utilisé pourgénérer un décompositeur analytique qui peut chercher à travers l'index.

— Org.apache.lucene.search : il se charge de fournir les objets pour chercher dans les indexes. Il fourni les classes suivantes :

- IndexSearcher : la classe IndexSearcher est la classe qui se charge de l'ouverture de l'index en lecture seulement.
- Query : c'est la méthode la plus basique d'interrogation de lucene, elleest utilisée pour égaliser les documents qui contiennent des champs avecdes valeurs spécifiques.
- Hits : la classe Hits est un conteneur d'index pour classer les résultatsde recherche de document. Pour des raisons de pertinences, les exemplesde classements ne chargent pas tous les documents de l'index pour unerequête donnée, mais seulement une partie d'entre eux.

— Org.apache.lucene.store : représente une couche d'abstraction d'entrée sortie.

On y trouve les classes :

- Directory : Les fichiers peuvent être écrits une fois, lorsqu'ils sont créés.

Une fois qu'un fichier est créé, il ne peut être ouvert qu'en lecture ou

supprimé. L'accès aléatoire est autorisé à la fois en lecture et en écriture.

- FSDirectory : c'est une classe qui étend de la classe Directory, elle sert à

stocker des fichiers d'index dans le système de fichiers.

- Org.apache.lucene.util : les classes sont utilisées dans les autres paquetages.

Par exemple on y trouve la classe Version qui permet de préciser la version de lucene utilisée.

III.3.4.2. La recherche sous lucene :

Avant de soumettre une requête à lucene, il faut d'abord passer par la phase d'indexation. Le schéma suivant illustre le processus d'indexation et les classes utilisées.

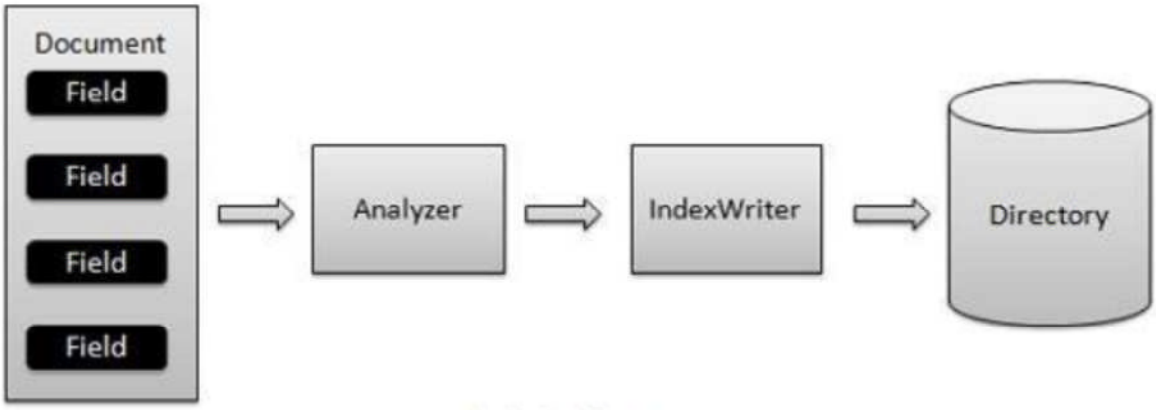


Figure III.10 : Processus d'indexation

La figure 3.3 met en oeuvre deux principales classes : Analyzer et IndexWriter.

Le fonctionnement est le suivant :

Nous ajoutons le ou les documents contenant le ou les champs à IndexWriter qui analyse le ou les documents à l'aide de l' Analyzer , puis on crée, ouvre ou édite les index selon les besoins et on les stocke ou met à jour dans un répertoire (Directory).

__ IndexWriter est utilisé pour mettre à jour ou créer des index et c'est la classe la plus importante du processus d'indexation.

Une fois l'index crée, nous pouvons effectuer une recherche sur une requête.

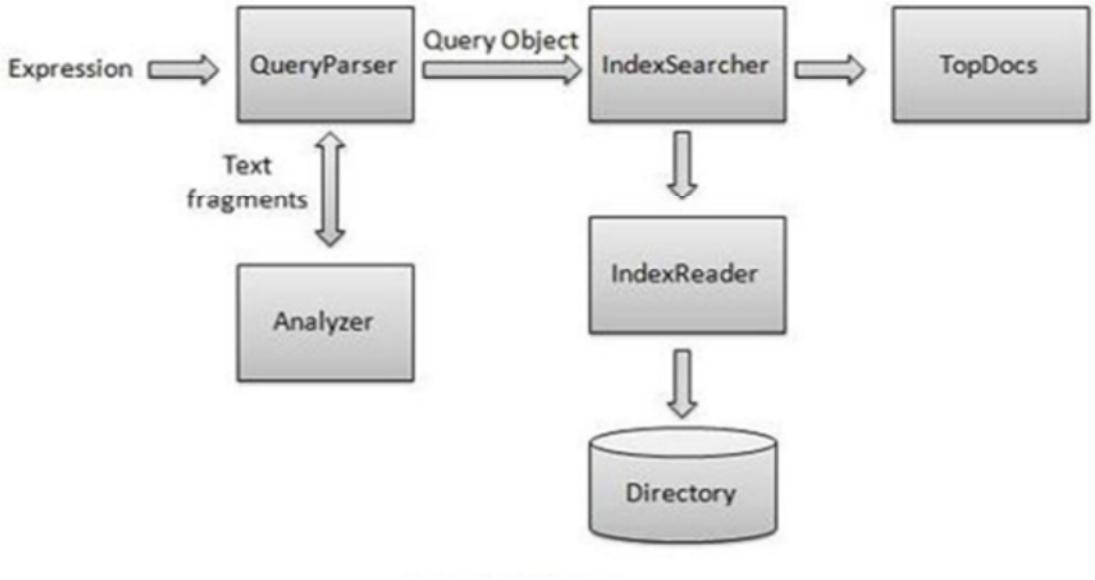


Figure III.11 : Processus de recherche

Le processus de recherche met en oeuvre les classes des packages org.apache.lucene.search et org.apache.lucene.queryparser :

— IndexSearcher : c'est la classe qui donne accès aux indexes en recherche.

— Analyzer : fait partie du processus de recherche pour normaliser les critères de recherche.

— QueryParser : analyseur de requêtes.

— Query : représente la requête de l'utilisateur et elle est utilisée par un index-Searcher.

— Hits : une collection d'éléments résultats de la recherche.

— Hit : un élément de la collection des résultats. Document : c'est l'unité contenant l'information.

III.4. Résultats des approches.

III.4.1 Tableau illustratifs des résultats.

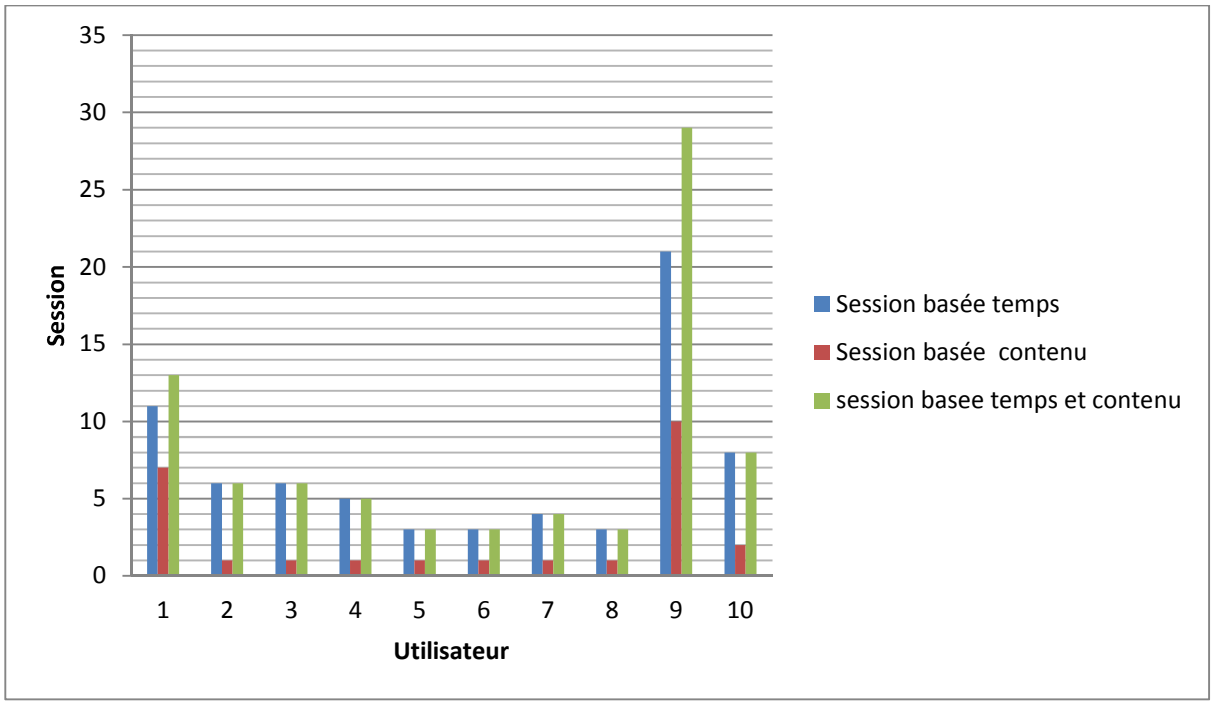
Nous présentons dans le tableau III.4 un récapitulatif des résultats obtenue selon les trois (03) approches (*Temps, Contenu, Temps et Contenu*).

USER	NBR Interactions	Session basée temps	Session basée contenu	session basée Temps et contenu
1	9	11	7	**13**
2	4	6	1	6
3	3	6	1	6
4	3	5	1	5
5	3	3	1	3
6	3	3	1	3
7	4	4	1	4
8	4	3	1	3
9	16	21	10	**29**
10	10	8	2	8

Tableau III.4 Résultats finaux des approches.

Nous présentons dans ce qui suit un graphe illustratifs des résultats comparatifs des différentes approches.

III.4.2 Graphe des résultats



Graphe 1 : Graphe des résultats

III.4.3 Interprétation des résultats du graphe :

D'après le graphe ci-dessus, on remarque que le nombre de session obtenu selon l'approche basée contenu est toujours inferieur au nombre de session obtenu selon les deux autre approches.

Par contre le nombre de session de recherche selon l'approche basée temps et contenu est toujo urs supérieur ou égale au nombre de session de recherche obtenu selon l'approche basée temps vu l'intégration d'un facteur qui est le centre d'intérêt ce qui augmente selon l'appartenance du profil de l'interaction à la même session de recherche.

On conclut que l'approche basée temps et contenu est l'approche la plus optimal pour délimiter une session de recherche.

III.5. Conclusion

Dans ce dernier chapitre, nous avons proposé le cadre expérimental de notre démarche Par la suite nous avons présenté un aperçu de notre implémentation, pour terminer nous avons montré les résultats de chaque approche pour toute la collection et nous avons déduit la meilleure approche de délimitation de session de recherche.

Conclusion

Dans le domaine de personnalisation de la recherche, une session de recherche regroupe l'ensemble des interactions effectuées par l'utilisateur. Elle est caractérisée par une durée et un centre d'intérêt lié au requête définissant ses interactions.

La problématique majeure est liée à la délimitation de session de recherche et l'évaluation sur une collection réelle.

Pour remédier a cette problématique nous avons évalué trois (03) méthodes de détection automatique de changement de session de recherche a savoir l'approche basée temps , basée contenu, et basée temps et contenu et nous les avons implémenté sur une collection réel .

D'après les résultats obtenus on a déduit que la meilleure approche est l'approche basée temps et contenu comparativement aux 2 autres approches.

Perspectives

Comme perspective nous envisageons.

- Elargir la collection de test réel pour une meilleure définition de profil et une meilleure estimation de la durée d'une session.
- La détection automatique des profils récurrents et persistants.

.

Bibliographie

[R. Baeza-Yates and R. A. Ribeiro-Neto, 99]

R. Baeza-Yates and R. A. Ribeiro-Neto. *Modern Information Retrieval*. New York :ACM Press ; Harlow England : Addison-Wesley, cop., 1999.

[N. Belkin and W. Croft,1992].

N. Belkin and W. Croft. Information filltering and information retrieval : Two sides of the same coin ? *Communication of the ACM*, 35(12) :29.38, 1992.

[C. Bradford and I. Marshall,99]

C. Bradford and I. Marshall. Analysing users www search behaviour. *Lost in the Web - Navigation on the Internet, IEE Colloquium*, 6(169) :1.4, 1999.

[M. Bates, 1981]

M. Bates. Search techniques. In *Annual Review of Information Science and Technology 16*, pages 139.169. M.E. Williams, ed., 1981.

[D. C. Blair and M. E. Maron,85]

D. C. Blair and M. E. Maron. An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Commun. ACM*, 28(3) :289.299, 1985.

[W. Croft et all, 95]

W. Croft, R. Cook, and D.Wilder. Providing government information on the internet : Experiences with thomas. In *Proceedings of the Second Annual Conference on the Theory and Practice of Digital Libraries (DL'95)*, pages 19.24, Austin, TX, Juin 1995.

[F. Crestani et all, 98]

F. Crestani, M. Lalmas, C. J. Van Rijsbergen, and I. Campbell. Is this document relevant ? ... probably. *A Survey of Probabilistic Models In Information Retrieval, ACM Computing Surveys*, 30(4), December 1998.

[B. Krovetz, 97]

B. Krovetz. Homonymy and polysemy in information retrieval. In P. R. Cohen and W. Wahlster, editors, *Proceedings of the e COLING/ACL'97*, pages 72.79, Somerset, New Jersey, 1997. Association for Computational Linguistics

[G. Miller , 95]

G. Miller. Wordnet : a lexical database for english. *Commun. ACM*, 38(11) :39.41, 1995.

[E. Voorhees, 94]

E. Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 61.69, Dublin, Ireland, 1994. Springer-Verlag New York, Inc.

[G. Brajnik et all, 96]

G. Brajnik, S. Mizzaro, , and C. Tasso. Evaluating user interfaces to information retrieval systems : a case study on user support. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 128.136, Zurich, 1996.

[K. Sparck Jones, *71]*

K. Sparck Jones. *Automatic Keyword Classi_cation for Information Retrieval*. Automatic keyword CL, London, 1971.

[Y. Qiu and H. Frei, 93]

Y. Qiu and H. Frei. Concept based query expansion. In *Proc. 16th Int'l ACM SIGIR Conf. R & D in Information Retrieval*, pages 160.169, 1993.

[V. Claveau and P. Sébillot , avril 2004.]

V. Claveau and P. Sébillot. Extension de requêtes par lien sémantique nom-verbe acquis sur corpus. In *Actes de la 11ème conférence de Traitement automatique des langues naturelles*, Fês, Maroc, avril 2004.

[Y. Jing and W. Croft, 1994.]

Y. Jing and W. Croft. An association thesaurus for information retrieval. In *Proceedingsof the 4th International Conference Recherche d'Information Assistee par Ordinateur*, pages 146.160, New York, US, 1994.

[S. Deerwster et all, 1990]

S. Deerwster, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society Information Science and Technology*, 41(6) :391.407, 1990.

[S. Robertson et all, 2000.]

S. Robertson, S. Walker, and M. Beaulieu. Experimentation as a way of life : Okapi at trec. *Information Processing & Management*, 36(1) :95.108, 2000.

[L. Tamineet all, 2003.]

L. Tamine, C. Chrisment, and M. Boughanem. Multiple query evaluation based on an enchanced genetic algorithm. *Information Processing & Management*, 39(2) :215. 231, 2003.

[J. Rocchio, 71]

J. Rocchio. Relevance feedback in information retrieval. In *The SMART retrieval system - experiments in automatic document processing*, pages 313.323, Englewood Cliffs, 1971. Prentice Hall.

[K. Kwok,1989].

K. Kwok. A neural network for probabilistic information retrieval. *SIGIR Forum*, 23(SI) :21.30, 1989.

[M. Boughanem and C. Soulé-Dupuy, 97]

M. Boughanem and C. Soulé-Dupuy. Mercure at trec-6. In *6th International Conference on Text Retrieval*, pages 321.328, Washington, USA, novembre 1997. E. M.Voorhees and D. K. Harman Editors.

[M. Boughanem et all, 99]

M. Boughanem, C. Chrisment, and C. Soulé-Dupuy. Query modi_cation based on relevance back-propagation in an ad hoc environment. *Information processing & management*, 35(2) :121.139, 1999.

[L. Tamine, 00]

L. Tamine. *Optimisation de requêtes dans un système de recherche d'information : approche basée sur l'exploitation de techniques avancées de l'algorithmique génétique*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France, décembre 2000.

[M. Boughanem et all 2002]

M. Boughanem, C. Chrisment, and L. Tamine. On using genetic algorithms for multimodal relevance optimisation in information retrieval. *Journal of American Society in Information Systems*, 53(11) :934.942, 2002.

[S. Robertson and K. Sparck Jones, 76]

S. Robertson and K. Sparck Jones. Relevance weighting for search terms. *Journal of The American Society for Information Science*, 27(3) :129.146, 1976.

[I. Ruthven and M. Lalmas, 03]
I. Ruthven and M. Lalmas. A survey on the use of relevance feedback for information access systems. *Knowledge Engeneering review*, 18(2) :95.145, 2003.

[R. W. White et all 03]
R. W. White, I. Ruthven, and J. M. Jose. A study of factors affecting the utility of implicit relevance feedback. In *Proceedings of the 28 th annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 15. 19. Marchionini, G. Moffat, A Tait, J Baeza-Yates, R Ziviani, N Eds, August 2003.

[P. Vakkari, 00].
P. Vakkari. Relevance and contributing information types of searched documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 2.9, Athens, Greece, 2000. ACM Press.

[L. B. Lorigo, H. Pan, T. Hembrooke, 06;]
L. B. Lorigo, H. Pan, T. Hembrooke, L. Joachims, and G. Granka. The in_uence of task and gender on search and evaluation behavior using google. *Information Processing & Management*, 42 :1123.1131, 2006.

[I. Kang and G. Kim, 04],
I. Kang and G. Kim. Integration of multiple evidences based on a query type for web search. *Information Processing & Management*, 40(3) :459.478, 2004.

[P. Vakkari, 01].
P. Vakkari. A theory of the task-based information retrieval process : a summary and generalisation of a longitudinal study. *Journal of documentation*, 57(1) :44.60, 2001.

[S. Yu, D. Cai, J. Wen, and W. Ma, 03].

S. Yu, D. Cai, J. Wen, and W. Ma. Improving pseudo-relevance feedback in web information retrieval using web page segmentation. In *Proceedings of the 12th international conference on World Wide Web*, pages 11.18, Budapest, Hungary, 2003. ACM.

[C. Zhai and J. Cohen, 03].
C. Zhai and J. Cohen. Beyond independent relevance : Methods and evaluation metrics for subtopical retrieval. In *Proceedings of the 27th annual international ACM SIGIR Conference on Research and development in Information retrieval*, pages 10. 17, August 2003.

[N. Belkinet et all, 01];

N. Belkin, C. Cool, D. Kelly, S.-J. Lin, S. Y. Park, J. Perez-Carballo, and C. Sikora.

Iterative exploration, design and evaluation of support of query reformulation in interactive information retrieval. *Information Processing & Management*, 37(3), 2001.

[R. W. White et all, 03]
R. W. White, I. Ruthven, and J. M. Jose. A study of factors affecting the utility of implicit relevance feedback. In *Proceedings of the 28 th annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 15. 19. Marchionini, G. Moffat, A Tait, J Baeza-Yates, R Ziviani, N Eds, August 2003.

[J. Janes, 91].
J. Janes. Relevance judgements and the incremental presentation of document representation. *Information Processing & Management*, 27(6) :629.646, 1991.

[X. Shen, B. Tan, and C. Zhai, 05].

X. Shen, B. Tan, and C. Zhai. Context-sensitive information retrieval using implicit feedback. In *In Proceedings of the 29th annual internationsl ACM SIGIR Conference on Research and development in Information retrieval*, pages 43.50, Salvador, Brazil, 2005.

[Baziz, 05]

M. BAZIZ, « *Indexation conceptuelle guidée par ontologie pour la recherche d'information* ». Thèse de doctorat de l'université de Paul Sabatier, spécialité informatique, 2005.

[Tamine, 05] L.LechaniTamine, M.Boughanem, « *Accès personnalisée a l'information : approches et techniques* », IRIT : institut de recherche en informatique de Toulouse, Equipe SIG/RFI, Rapport interne, Janvier 2005.

[Tamine, 06] L. Tamine-Lechani, M. Boughanem, C. Chrisment, « *Accès personnalisé à l'information: Vers un modèle basé sur les diagrammes d'influence* », Institut de Recherche en Informatique de Toulouse, Equipe Systèmes d'Information Généralisés, 2006.

[Boughanem 01] M. BOUGHANEM, M. TMAR, « Filtrage d'information par combinaison d'un profil positif et profil négatif », IRIT/SIG, Compus Univ Toulouse III, Université de Nantes-Paris X, 3° congrès du chapitre français di l'ISKO, juillet 2001, p209-217.

[Kostadinov, 04] D.Kostadinov, « *personnalisation de l'information et gestion des Profils utilisateurs* », laboratoire PRiSM, université de Versailles,Article, France 2004.

[kostadinov, 04] kostadinov D, Bouzeghoub M, « *Une approche multidimensionnelle pour la personnalisation de l'information* », INRIA rocquencourt et laboratoire PRISM, université de Versailles, France, 2004.

[Kien, 06] Kien D N, « *Moteur de composition pour le système d'information sémantique et adaptatif* », mémoire de fin d'études master en informatique, l'institue de la francophonie pour l'informatique, 13 septembre 2006.

[KOS, 04] D. KOSTADINOV, M. BOUZEGHOUB, *« Une approche multidimensionnelle pour la personnalisation de l'information ».* INRIA Rocquencourt et Laboratoire PRiSM, Université de Versailles, France, 2004.

[Dahak, 06] F. DAHAK, « Indexation des documents semi-structurés: Proposition d'une approche basée sur le fichier inversé et le Tree ». Thèse de Magister, 2005.

[Daoud, 09] M.DAOUD, « Accès personnalisé à l'information : approche basée sur l'utilisation d'un profil dérivé d'une ontologie de domaines à travers l'historique des sessions de recherche » thèse de doctorat de l'université Paul Sabatier,

Université Paul Sabatier, Toulouse, 2009.

[Gaussier, 03] Gaussier E, Jacquemin C, Zweignebaum P, « Traitement automatique des langages et recherche d'information ». Chapitre 2, p72:96, « Assistance intelligente à la recherche d'information », Hermes science, 2003.

[Gowan, 03] J.P McGowan, « A multiple model approach to personalized information access », Thesis of master in computer science, Faculty of science, university College Dublin, February 2003.

[Horvitz, 98] Horvitz E, Breese J, Heckerman, D. Hovel D, Rommelse K, « The lumiere project: Bayesian user modeling for inferring the goals and needs of software users», Fourteenth conference on uncertainty in artificial intelligence, Madison, Wisconsin, 256:265, 1998.

[Jenningd, 93] Jenningd, A.Higuchi, H, « A user model neural for a personal news service user modeling and user adapted interaction ». 1993.

[Kien, 06] Kien D N, « Moteur de composition pour le système d'information sémantique et adaptatif », mémoire de fin d'études master en