Ministère de l'enseignement supérieur et de la recherche scientifique

Université SAAD DAHLEB Blida

Mémoire

En vue de l'obtention du diplôme de Master

En Informatique

Option:

Ingénierie de logiciel

IL

Présenté par:

ZEGAOUI Nasreddine

SID Mohamed Fares

Thème

Réinjection automatique de la pertinence dans une recherche d'information contextuelle

Encadré par : M[r] FERFERA Soufiane

2012-2013

Soutenu publiquement le 10/09/2013

Remerciements

Nous tenons à exprimer notre profonde gratitude à Monsieur FERFERA Soufiane, pour avoir diriger ce mémoire dans la continuité de notre stage de Master. Son encadrement, ses critiques constructives, ses précieux conseils, ses relectures acharnées de nos travaux nous ont été d'une aide précieuse. Pour tout cela, sa confiance et sa disponibilité du début à la fin du mémoire, je le remercie vivement et qu'il trouve ici l'expression de ma considération profonde.

Nous tenons à remercier les membres du jury qui ont eu l'amabilité de bien vouloir participer à ce jury, et l'intérêt qu'ils ont porté à ce travail.

Nous souhaitons aussi remercier toutes les personnes qui ont contribué de prés ou de loin à l'accomplissement de ce travail.

Résumé

Il devient de plus en plus difficile de trouver des informations sur le web qui satisfont les besoins des utilisateurs. Le problème n'est pas tant la disponibilité de l'information mais sa pertinence relativement à un contexte d'utilisation particulier. Dans ce cadre, la recherche d'information contextuelle émerge comme un domaine à part entière. La prise en compte des informations du contexte de l'utilisateur constitue une piste intéressante pour améliorer les résultats de la recherche. L'idée est d'exploiter les données du contexte de recherche pour sélectionner l'information pertinente en réponse à une requête utilisateur qui est souvent insuffisante pour permettre la sélection de documents pertinents répondant au besoin de l'utilisateur. Cette information peut être utilisée dans la requête initiale de l'utilisateur afin d'améliorer les résultats de recherche (réinjection automatique de pertinence).

Mots clés : Recherche d'information, recherche d'information contextuelle, système de recherche d'information, reformulation de requête, réinjection automatique de pertinence.

Abstract

It is becoming more difficult to find information on the web that meets the needs of users. The problem is not so much the availability of information, but its relevance to a particular context of use. In this context, the search for contextual information emerges as a field in itself. Taking into account the context information of the user is an interesting way to improve search results. The idea is to use the data in the context of research to select relevant information in response to a user query that is often insufficient to allow the selection of relevant documents to meet user needs. This information can be used in the initial request from the user to improve search results (automatic relevance feedback).

Keywords: Information search, contextual information retrieval, system information retrieval, query reformulation, automatic relevance feedback.

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations

GPS: Global Positioning System
H-M: Homme-Machine
HTML: Hyper Text Markup Language
IAu: Indexation Automatique
IC: Indexation Contrôlée
IE: Internet Explorer
IL: Indexation Libre
IM: Indexation Manuelle
ISAu: Indexation Semi-Automatique
NASA: l'Académie des Sciences Nationale des Etats-Unis d'Amérique.
PDA: Personal Digital Assistant
PPV: Personalized PageRank Vector
PRF: Pseudo Relevance Feedback
RDN: Resource Discovery Network
RF: Relevance Feedback
RFA : Relevance Feedback Automatic
RI : Recherche d'Information
RIC : Recherche d'Information Contextuelle
RSI: Recherche Sociale d'Information
RSV: Retrieval Status Value
SRI: Système de Recherche d'Information
SRIP: Système de RI Personnalisé
SRSI: Système de Recherche Sociale d'Information
TALN: Traitement Automatique du Langage Naturel
TG: Thesaurus Générique
TS: Thesaurus Spécifique
URL: Uniform Resource Locator
Vlib: Virtual Library
XML: Extensible Markup Language

Introduction générale :

Chercher une information sur le web devient un geste quotidien que font des utilisateurs diversifiés en âge, culture, spécialité et ayant des domaines d'intérêt variés. De nos jours, la richesse documentaire augmente et c'est essentiellement grâce à la croissance massive des documents numériques, souvent hétérogènes dans leur forme et leur contenu.

En raison de la diversité des masses d'informations, l'utilisateur a en général de plus en plus de difficultés pour accéder aux informations qui répondent à son besoin. L'élaboration de systèmes automatisés pour gérer ces masses d'informations est devenue dans un tel contexte une nécessité. La RI, domaine déjà ancien, est une branche en informatique qui s'intéresse à l'acquisition, l'organisation, le stockage et la recherche des informations. Elle propose des outils, appelés systèmes de recherche d'information (SRI), dont l'objectif est de capitaliser un volume important d'information et d'offrir des moyens permettant de localiser les informations pertinentes relatives à un besoin en information d'un utilisateur exprimé à travers une requête [5]. Les études montrent que l'approche généraliste des outils disponibles de RI sur le web qui répond invariablement les utilisateurs en renvoyant une même liste de résultats pour deux utilisateurs ayant émis la même requête et ayant pourtant des besoins en informations et des préférences de recherche différentes, est à l'origine des problèmes évoqués [1].

Les premières tentatives proposées permettant de pallier cette problématique s'apparentent à la RI adaptative. Le but de la RI adaptative est d'adapter le processus de RI au besoin précis de l'utilisateur en termes des documents pertinents associés. Les techniques dérivées de ce cadre concernent la reformulation de requête par réinjection de pertinence, l'expansion des requêtes par désambiguïsation ou le regroupement thématique des résultats. Malgré l'efficacité de la recherche obtenue par ces techniques, elles sont généralement limitées par la rétroaction explicite de la part des utilisateurs. En outre leur efficacité est relativement dépendante du niveau de familiarité de l'utilisateur avec le sujet de recherche et ne seront efficaces qu'après plusieurs itérations de recherche.

Suite aux limitations des techniques de la RI adaptative et dans le but d'améliorer la performance des SRI, plusieurs études ont été menées dans le but de mieux cerner la notion de pertinence du point de vue de l'utilisateur et d'identifier les différents facteurs ayant un impact sur cette notion et par conséquent sur la performance des SRI. Ces études confirment que la pertinence n'est pas une relation isolée entre un document et une requête, elle est définie en fonction du contexte dans lequel la recherche est effectuée. C'est ainsi qu'une nouvelle direction de recherche basée sur la RI contextuelle est apparue. L'objectif principal des travaux en recherche d'information contextuelle (RIC) est d'optimiser la pertinence des résultats de recherche, en impliquant deux étapes complémentaires : définition du contexte du besoin en information de l'utilisateur, puis adaptation de la recherche en le prenant en considération dans le processus de sélection de l'information.

La recherche d'information est un processus qui se base essentiellement sur la requête exprimée par l'utilisateur pour répondre à ses besoins. En effet, quel que soit le système de recherche utilisé, le résultat d'une recherche ne peut être pertinent si la requête ne décrit pas explicitement et clairement les besoins de l'utilisateur. Or, il est généralement reconnu que l'utilisateur se contente de donner quelques mots clés. Ces derniers sont issus d'une connaissance générale sur le sujet recherché. Par conséquent, les documents renvoyés par le système de recherche peuvent ne pas satisfaire les besoins de l'utilisateur.

La reformulation de requêtes est une des stratégies qui permet d'améliorer la construction d'une requête. Elle consiste de manière générale à enrichir la requête de l'utilisateur en ajoutant des termes permettant de mieux exprimer son besoin [6]. Une des techniques les plus répandues en RI est la reformulation par réinjection automatique de la pertinence. Elle consiste à extraire à partir d'un échantillon de documents les mots clés les plus pertinents et les ajouter à la requête.

Plusieurs outils et applications ont été développés en recherche d'information contextuelle et en recherche d'information par reformulation de requêtes. Il serait peut être intéressant de voir que peut donner la réalisation d'outils utilisant le réinjection de pertinence dans le contexte d'une recherche d'information contextuelle par rapport à une recherche d'information classique.

Nous nous intéressons donc dans le cadre de ce mémoire à la réinjection automatique de pertinence dans une recherche d'information contextuelle. Plusieurs questions se posent dans ce contexte, elles portent en général sur le choix d'un échantillon de documents et le choix de la notion du contexte. Plus précisément :

- En RI classique le choix d'un échantillon de documents est aléatoire en général en prenant les « k » premiers documents retournés par la requête initiale, et cela ne peut être bénéfique que si ces « k » premiers documents sont pertinents sinon on va avoir le problème de la dérivé de requête.
- Comment peut-on choisir l'échantillon de documents pertinents selon le contexte de l'utilisateur sans l'intervention de ce dernier pour garder l'idée général de la réinjection automatique de la pertinence ?
- Comment modéliser le contexte de l'utilisateur ?
- Comment intégrer et exploiter le contexte de l'utilisateur dans le processus de recherche pour trouver les documents pertinents selon ce dernier ?
- Comment extraire à partir des documents jugés pertinents les termes à ajoutée à la requête initiale ?

Afin de répondre aux questions listées précédemment, nous avons proposé un mécanisme de reformulation partant de la sélection d'échantillon de documents jugés pertinents selon le contexte de l'utilisateur jusqu'au renvoi d'un ensemble de documents répondant à la requête reformulée par la réinjection automatique de la pertinence.

Pour ce faire notre mémoire est divisé en deux principales parties : état de l'art et contribution.
Les deux principales parties sont organisées comme suit :

- La première partie, composée de trois chapitres présente un état de l'art. Le premier chapitre présente les concepts de base de la RI. Nous commençons par donner une définition de la RI et nous décrivons les différents modèles servant de cadre théorique

pour la modélisation du processus de RI. Nous illustrons également le processus de RI en présentant les étapes d'indexation, d'interrogation et de mise en correspondance. Nous parlons aussi sur les techniques de base de la RI adaptative. Par la suite nous présentons les outils de RI sur le web, et l'architecture générale des moteurs de recherche.

Le deuxième chapitre traite la RI contextuelle. Nous abordons les différentes définitions du contexte, les différentes taxonomies du contexte qu'elles ont apparu dans la littérature, les possibilités de son utilisation en RI, et nous passons par les systèmes de RIC. Nous parlons par la suite sur les approches développées en RI contextuelle et l'exploitation des dimensions du contexte dans la RI. Nous commençons par la RI contextuelle guidée par le profil utilisateur, et nous passons sur la RI contextuelle guidée par le contexte social, ensuite nous parlons sur la RI contextuelle guidée par le contexte mobile.

Le troisième chapitre développe la reformulation de requêtes et la réinjection automatique de pertinence. Nous commençons par les outils de base utilisés pour la reformulation et l'exploitation du contexte globale et local dans la reformulation automatique et ensuite nous parlons sur le processus général de la réinjection de pertinence et les approches à utiliser dans ce contexte. Ensuite nous parlons sur la réinjection automatique de pertinence comme un modèle de reformulation de requête, nous présentons le principe général de celle-ci, les formes, les modèles, les étapes de l'extraction des termes dans le corpus et des exemples de l'approche à utiliser dans la pondération des termes normalisés.

- La deuxième partie composée de deux chapitres présente notre travail et son évaluation. Un chapitre présente notre travail. Nous commençons par le choix de dimension de contexte choisi et son exploitation dans notre travail par une première reformulation pour trouver l'échantillon des documents pertinents, ensuite nous passons par la réinjection automatique de pertinence et l'exploitation de l'échantillon de documents pertinents dans la deuxième reformulation pour renvoyer aux utilisateurs le résultat final.

  L'autre chapitre présente l'implémentation et l'évaluation de notre travail. Nous commençons par l'implémentation où on parle sur les outils utilisés et l'environnement de développement, ensuite nous expliquons le fonctionnement de notre système. Nous évaluons par la suite notre système.

Chapitre 01

Concepts de base de la RI

1.1 Introduction :

La recherche d'information (RI) est une discipline de recherche qui intègre des modèles et des techniques dont le but est de faciliter l'accès à l'information pertinente pour un utilisateur ayant un besoin en information. Le besoin en information de l'utilisateur est souvent formulé en langage naturel par une requête décrite par un ensemble de mots clés. Pour une requête utilisateur, un système de RI permet de retrouver un sous-ensemble de documents susceptibles d'être pertinents, à partir d'une collection de documents, en réponse à cette requête.
L'essor du web a remis la RI face aux nouveaux défis d'accès à l'information, à savoir retrouver une information pertinente dans un espace diversifié de taille considérable et qui répond au besoin en information spécifique de l'utilisateur.
La limite majeure de la plupart des modèles de recherche classiques est qu'ils retournent la même liste des résultats pour une même requête soumise par des utilisateurs étant dans des contextes et/ou des situations de recherche différents et par conséquent ayant des besoins en information différents.
Les études [1] montrent que l'origine de ces limites réside en partie dans le fait que ces modèles sont basés sur une approche généraliste qui considère que le besoin en information de l'utilisateur est complètement représenté par sa requête, et délivrent alors des résultats en ne tenant compte que des critères de sélection par contenu et de la disponibilité des sources d'information.
Les premières techniques développées en RI dans le but de remédier à cette problématique s'apparentent à la RI adaptative. Il s'agit du développement de techniques de reformulation de requêtes, de désambigüisation du besoin derrière les requêtes ou du regroupement thématique des résultats de la recherche [2].

Ce chapitre est organisé en quatre grandes parties : la première présente les concepts de base de la RI et les différents modèles qui ont été proposés pour fournir un cadre théorique pour la modélisation du processus de RI. La deuxième partie décrit le processus de RI, à savoir les étapes d'indexation, d'interrogation et de mise en correspondance. La troisième partie décrit l'évolution de la RI classique à la RI adaptative. La quatrième partie sera consacrée à la RI sur le web en présentant les outils de recherche d'informations sur le web. Le chapitre sera terminé par une conclusion.

1.2. La recherche d'information :

Plusieurs définitions de la recherche d'information ont vu le jour dans ces dernières années, nous citons dans ce contexte les trois définitions suivantes :

1.2.1. Définitions :

- *Définition 1* : 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. [2]

- *Définition 2* : La recherche d'information est une discipline de recherche qui intègre des modèles et des techniques dont le but est de faciliter l'accès à l'information pertinente pour un utilisateur ayant un besoin en information. [3]

- *Définition 3* : la recherche d'information est l'ensemble des méthodes, procédures et techniques permettant de sélectionner l'information dans un ou plusieurs fonds de documents plus ou moins structurés. [4]

Toutes ces définition partagent l'idée que la RI à pour objet d'extraire d'un document ou d'un ensemble de documents les informations pertinentes qui reflètent un besoin d'information.

1.2.2. Concepts de base de la RI :

– *Document :* On appelle document toute unité d'information qui peut constituer une réponse à un besoin en information/requête d'un utilisateur. Un document peut être un texte, un morceau de texte, une image, une bande vidéo, etc. [4]

– *Requête :* Une requête est une formulation du besoin d'information d'un utilisateur. Elle peut être vue comme étant une description sommaire des documents ciblés par la recherche. Pour une recherche documentaire donnée, l'utilisateur doit soumettre une requête au moteur de recherche dans laquelle il spécifie les mots clés représentant son besoin en information. [4]

– *Pertinence* **:** En RI indique dans quelle mesure les documents retournés par le système de RI répondent au besoin d'information de l'utilisateur. Cette notion représente un critère majeur de l'évaluation des performances du système de RI. La pertinence, qui est l'objet principal de tout système de RI, constitue une notion fondamentale en RI. Elle peut être définie comme la correspondance entre un document et une requête ou encore une mesure d'informativité du document à la requête. [4]

–*Modèle de recherche* **:** Il représente le modèle du noyau d'un SRI. Il comprend la fonction de décision fondamentale qui permet d'associer à une requête, l'ensemble des documents pertinents à restituer. Il est utilisé pour la recherche d'informations proprement dite et est étroitement lié au modèle de représentation des documents et des requêtes. [3]

1.2.3. Les modèles de RI :

Les travaux de recherche dans le domaine de la RI ont conduit à la proposition de nombreux modèles de RI [5].

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.

1.2.3.1. Modèle booléen :

Le modèle booléen est historiquement le premier modèle de RI, et est basé sur la théorie des ensembles. Un document est représenté par une liste de termes (termes d'indexation). Une requête est représentée sous forme d'une équation logique. Les termes d'indexation sont reliés par des connecteurs logiques ET, OU et NON.

Le processus de recherche mis en œuvre consiste à effectuer des opérations sur l'ensemble de documents afin de réaliser un appariement exact avec l'équation de la requête. L'appariement exact est basé sur la présence ou l'absence des termes de la requête dans les documents.

La décision binaire sur laquelle est basée la sélection d'un document ne permet pas d'ordonner les documents renvoyés à l'utilisateur selon un degré de pertinence [6].

1.2.3.2. Modèle vectoriel :

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

Chaque document est représenté par un vecteur : $D_j = (d_{1j}, d_{2j}, \ldots, d_{ij}, \ldots, d_{Tj})$

Chaque requête est représentée par un vecteur : $Q = (q_1, q_2, \ldots, q_i, \ldots, q_T)$

Avec : d_{ij} Poids du terme t_i dans le document D_j

q_i Poids du terme t_i dans la requête Q

La fonction de calcul du coefficient de similarité entre chaque document, représenté par le vecteur D_j $(d_{1j}, d_{2j}, \ldots, d_{Tj})$ et la requête, représentée par le vecteur $(q_1, q_2, \ldots, q_i, \ldots, q_T)$ est appelée *Retrieval Status Value* ou *RSV*.

Ce coefficient de similarité est calculé sur la base d'une fonction qui mesure la colinéarité des vecteurs : document et requête.

La similarité entre deux textes (requêtes ou documents) dépend ainsi des poids des termes coïncidant dans les deux textes. Il est donc possible de classer les documents par ordre de pertinence décroissante [7].

1.2.3.3. Modèle probabiliste :

Ce modèle aborde le problème de la recherche d'information dans un cadre probabiliste. La pertinence document-requête est traduite par le calcul de la probabilité de pertinence d'un document par rapport à une requête. La pertinence entre un document et une requête est mesurée par le rapport entre la probabilité qu'un document *D* donné soit pertinent pour une requête *Q*, notée $p(R/D)$, et la probabilité qu'il soit non pertinent, notée p (nonp/*D*), ou *R* est l'élément de pertinence et (non*R)* de non pertinence. Ces probabilités sont estimées par les probabilités conditionnelles selon qu'un terme de la requête est présent, dans un document pertinent ou dans un document non pertinent [5].

1.3. Système de recherche d'information :

1.3.1. Définition :

- *Définition 1* : Un système de recherche d'information est défini par un langage de représentation des documents (qui peut s'appliquer à différents corpus de documents) et des requêtes qui expriment un besoin de l'utilisateur (sous forme de mots-clés par exemple) et une fonction de mise en correspondance du besoin de l'utilisateur et du corpus de documents en vue de fournir comme résultats des documents pertinents pour l'utilisateur, c'est-à-dire répondant à son besoin d'information. [3]

- *Définition 2* : Un système de recherche d'information est un système qui permet de retrouver, à partir d'une collection de documents, les documents susceptibles d'être pertinents à un besoin en information d'un utilisateur exprimé sous forme d'une requête. [2]

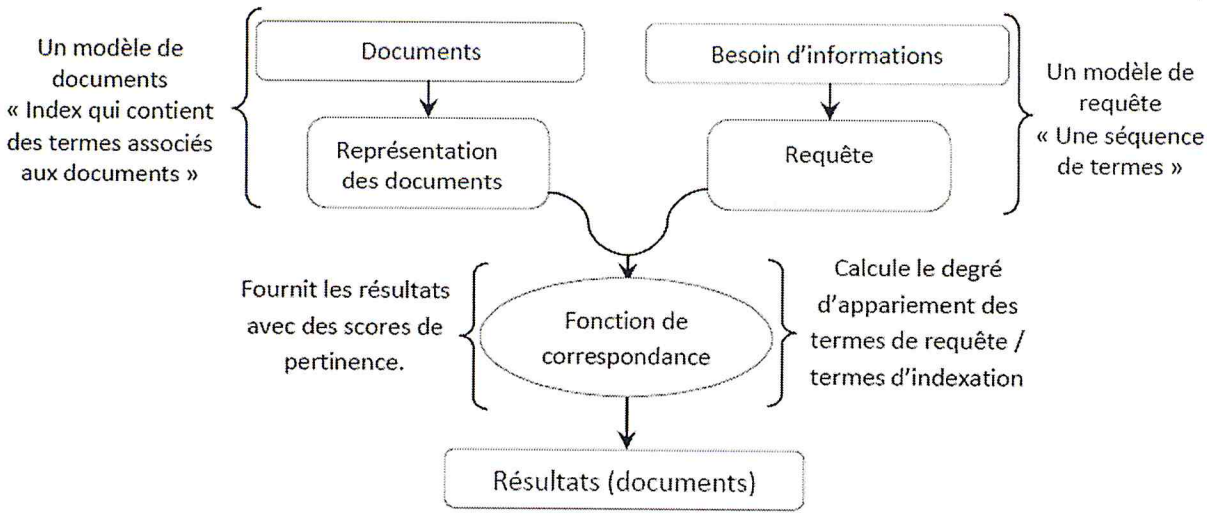

Figure 1: Système de Recherche d'Information [3]

1.3.2. Processus de recherche d'information :

L'objectif fondamental d'un processus de RI est de sélectionner les documents "*les plus proches*" du besoin en information de l'utilisateur décrit par une requête. Pour cela, le système de recherche regroupe un ensemble de méthodes et procédures permettant la gestion des collections de documents stockés sous forme d'une représentation intermédiaire permettant de refléter aussi fidèlement que possible leurs contenus sémantiques.

1.3.2.1. Indexation :

L'indexation consiste à déterminer et à extraire les termes représentatifs du contenu d'un document. Le résultat de l'indexation constitue ce que l'on nomme le descripteur du document. 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 [5]. L'indexation est une étape très importante dans la recherche d'information car sa qualité dépend de la qualité des réponses du système et donc les performances de ce dernier. Une bonne indexation doit permettre de retrouver tous les documents pertinents au besoin de l'utilisateur et pas (ou peu) de documents non pertinents pour celui-ci [8]. On distingue deux types d'indexation, indexation libre et indexation contrôlée [9] :

– *Indexation libre(IL)* : l'indexeur extrait les mots-clés d'un document ou les choisit librement sans l'aide de sources de connaissance. L'indexation peut être améliorée en filtrant les mots fonctionnels pour éliminer les non-descripteurs du document [4].

– *Indexation contrôlée(IC)* : le langage d'indexation est construit à partir d'un ensemble de termes préalablement définis et organisés généralement dans un thésaurus. Lorsqu'un document est analysé, on ne garde que les mots clés qui appartiennent à ce thésaurus [9].

Notons que chacun de ces deux types peut être manuelle, automatique ou semi-automatique [7] :

– *Indexation manuelle(IM)* : Lors de l'indexation manuelle, un expert dans le domaine choisit les termes qu'il juge pertinents dans la description du contenu sémantique du

document. Ce type d'indexation permet d'avoir un vocabulaire d'index contrôlé ce qui permet d'accroître la consistance et la qualité de la représentation obtenue. Toutefois cette approche est subjective d'une part car elle dépend des connaissances de l'opérateur et d'autre part inapplicable pour une collection volumineuse [2].

– *Indexation automatique(IAu)* : Ce type d'indexation ne fait pas intervenir l'expert. L'indexation automatique repose sur des algorithmes associant automatiquement des descripteurs à des parties de document. Dans le cas des documents textuels, chaque mot est potentiellement un index du document qui le contient [2].

– *Indexation semi-automatique(ISAu)* : C'est une combinaison des deux méthodes précédentes : un premier processus automatique permet d'extraire les termes du document. Cependant, le choix final des descripteurs est laissé au spécialiste du domaine, qui utilise un vocabulaire contrôlé sous forme de thésaurus ou de base terminologique [2].

1.3.2.2. Interrogation :

L'interrogation du système implique un processus d'interaction de l'utilisateur avec le SRI [2]. Cette interaction comprend :

– la formulation d'une requête par l'utilisateur traduisant son besoin en information.

– la représentation de la requête sous forme interne selon le langage d'indexation défini.

– la correspondance entre la requête et les documents par exploitation de l'index et la présentation des résultats.

Donc cette phase nécessite un modèle de représentation du besoin de l'utilisateur, appelé modèle de requêtes, ainsi qu'une fonction de correspondance qui doit évaluer la pertinence des documents par rapport à la requête. La réponse du système est un ensemble de références à des documents qui obtiennent une valeur de correspondance élevée. Cet ensemble est généralement présenté sous la forme d'une liste ordonnée suivant la valeur de correspondance. La satisfaction de l'utilisateur concernant les documents retournés par le système, n'est pas toujours acquise. Certains systèmes permettent aux utilisateurs de marquer parmi les documents résultats ceux qu'ils jugent pertinents ou non pertinents. Ces jugements sont alors pris en compte pour définir une nouvelle requête, il s'agit du processus de reformulation. Ce processus n'est pas toujours automatique, une stratégie classique d'utilisation des systèmes de recherche d'information consiste à reformuler manuellement la requête en tenant compte des documents pertinents et non pertinents obtenus [3].

1.3.2.3. Fonction de correspondance :

Tout système de recherche d'information s'appuie sur un modèle de recherche d'information. Ce modèle se base sur une fonction de correspondance qui met en relation les termes d'un document avec ceux d'une requête en établissant une relation d'égalité entre ces termes. Cette relation d'égalité représente la base de la fonction de correspondance et, par la même, du système de recherche d'information. Il existe un certain nombre de modèles théoriques dans la littérature les plus connus étant le « Modèle Booléen», le «Modèle Vectoriel», et le «Modèle Probabiliste ». Dans le modèle booléen, les requêtes sont représentés sous forme de termes reliés par des opérateurs booléens (ET, OU, NON, . . .). Le modèle vectoriel considère les documents et les requêtes comme des vecteurs pondérés, chaque élément du vecteur représentant le poids d'un terme dans la requête ou le document. Le modèle probabiliste tente d'estimer la probabilité qu'un document donné soit pertinent pour une requête donnée [3].

1.4. De la RI classique à la RI adaptative :

Le principe fondamental commun à tous les modèles classiques de RI suppose que les documents sélectionnés doivent contenir les mêmes mots que ceux formulés par l'utilisateur et que la requête représente ce besoin en information. 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. Lors de l'appariement requête/document, seuls les documents qui sont les plus proches sémantiquement du besoin de l'utilisateur sont sélectionnés. De ce fait, plus les termes d'indexation sont représentatifs du contenu sémantique des documents et de la requête, plus la pertinence des documents sélectionnés est améliorée.
Néanmoins, dans la pratique la majorité des requêtes exprimées par les utilisateurs sont courtes et ambiguës [5]. En outre, cette liste de termes ne correspond pas forcément à ceux utilisés pour indexer les documents pertinents de la collection et manque souvent de termes significatifs pouvant exprimer effectivement le besoin en information de l'utilisateur [5].
Ceci mène aux problèmes de disparité des termes et d'ambiguïté [5] en recherche d'information, l'utilisateur et l'auteur d'un document n'utilisent pas nécessairement le même vocabulaire. Ainsi, un document peut être pertinent même s'il ne contient pas les mêmes termes que ceux de la requête. Cependant, dans les SRI classiques, un tel document ne sera pas retourné à l'utilisateur à cause du défaut d'appariement document-requête. De plus, une même requête exprimée par deux utilisateurs ayant des besoins différents va être exécutée de façon similaire par le SRI, et aucune distinction ne sera apportée aux résultats de recherche. De ce fait, les performances d'un SRI, ne dépendent pas uniquement de l'efficacité et la qualité des mécanismes d'indexation et d'appariement, mais de façon non négligeable de la capacité du SRI de prendre en considération les besoins de l'utilisateur pour mieux répondre à leurs attentes. De ce constat est apparu un nouvel axe de recherche, celui de la RI adaptative [5].
La RI adaptative tente d'exploiter des informations additionnelles, au-delà de la requête, généralement extraites des interactions de l'utilisateur avec le SRI, dans le but d'améliorer la recherche. On peut distinguer trois principales classes de techniques développées en RI adaptative [1]:

- les techniques de reformulation de requête.
- les techniques de désambiguïsation du sens des mots de la requête.
- les techniques de regroupement thématique des résultats de recherche Grouper.

1.4.1. Reformulation de requêtes :

Il est souvent difficile, pour l'utilisateur, de formuler son besoin exact en information. Par conséquent, les résultats que lui fournit le SRI ne lui conviennent parfois pas. Retrouver des informations pertinentes en utilisant la seule requête initiale de l'utilisateur est toujours difficile et c'est à cause de l'imprécision de la requête. Afin de faire correspondre au mieux la pertinence utilisateur et la pertinence du système, une étape de reformulation de la requête est souvent utilisée. La requête initiale est traitée comme un essai (naïf) pour retrouver de l'information. Les documents initialement présentés sont examinés et une formulation améliorée de la requête est construite, dans l'espoir de retrouver plus de documents pertinents. [12]
Plusieurs approches de reformulation de requêtes ont été apparu pour objectif de construire une nouvelle requête par ajout, suppression et/ou repondération des termes. Une parmi ces

approches basée sur l'ajout 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, ou à partir de thésaurus [5]. Autre approche de reformulation est apparue, elle est nommé réinjection de pertinence et/ou non pertinence. L'idée est de présenter à l'utilisateur une liste de documents. Après les avoir examiné , l'utilisateur indique ceux qu'il considère pertinents, ensuite on sélectionne les termes importants appartenant aux documents jugés pertinents par l'utilisateur, et on renforce l'importance de ces termes dans la nouvelle formulation de la requête [12]. Face à cette dernière approche, il existe une autre qui garde la même idée mais dans celle-ci l'utilisateur n'intervient pas, et on considère que les *k* premiers documents sont pertinents (documents pseudo-pertinents), cette approche est appelée pseudo-réinjection de la pertinence (*Blind Feedback* ou encore *Pseudo Relevance Feedback*, notée PRF) [4].

Toutes ces approches nous allons les détailler dans le chapitre 4 qui traite la reformulation de requête.

1.4.2. Désambiguïsation du sens des mots de la requête :

Ces techniques consistent à aider l'utilisateur d'exprimer mieux son besoin en information et orienter la recherche vers les documents portant l'intention de recherche de l'utilisateur. Elles permettent à l'utilisateur de saisir le vrai sens évoqué par les termes de la requête et l'exploiter dans des techniques d'extension des langages de requêtes. La plupart de ces techniques se basent sur l'exploitation des interfaces de clarification interactives à base d'ontologie [2]. D'autres approches se basent sur la définition des paramètres mesurables à partir de la requête ou à partir du profil des premiers documents retournés par la requête. Dans ce sens, un nouveau type de réinjection de pertinence qui est le profil de requête est exploité afin de détecter des besoins en information divers et des critères de qualité de l'information derrière la requête, tels que le thème de recherche et l'information récente [1]. Une étude comporte la génération des profils temporels et thématiques des requêtes où le profil temporel de la requête est défini par la distribution des N premiers documents retournés pour la requête en fonction de leur date de création. Cette approche permet à l'utilisateur d'affiner sa requête en choisissant d'une part le sujet d'intérêt correspondant à la requête et d'autre part de choisir ses dates préférées identifiées à partir du profil temporel de la requête afin de générer une requête ciblant les résultats estampillés selon ces dates [2].

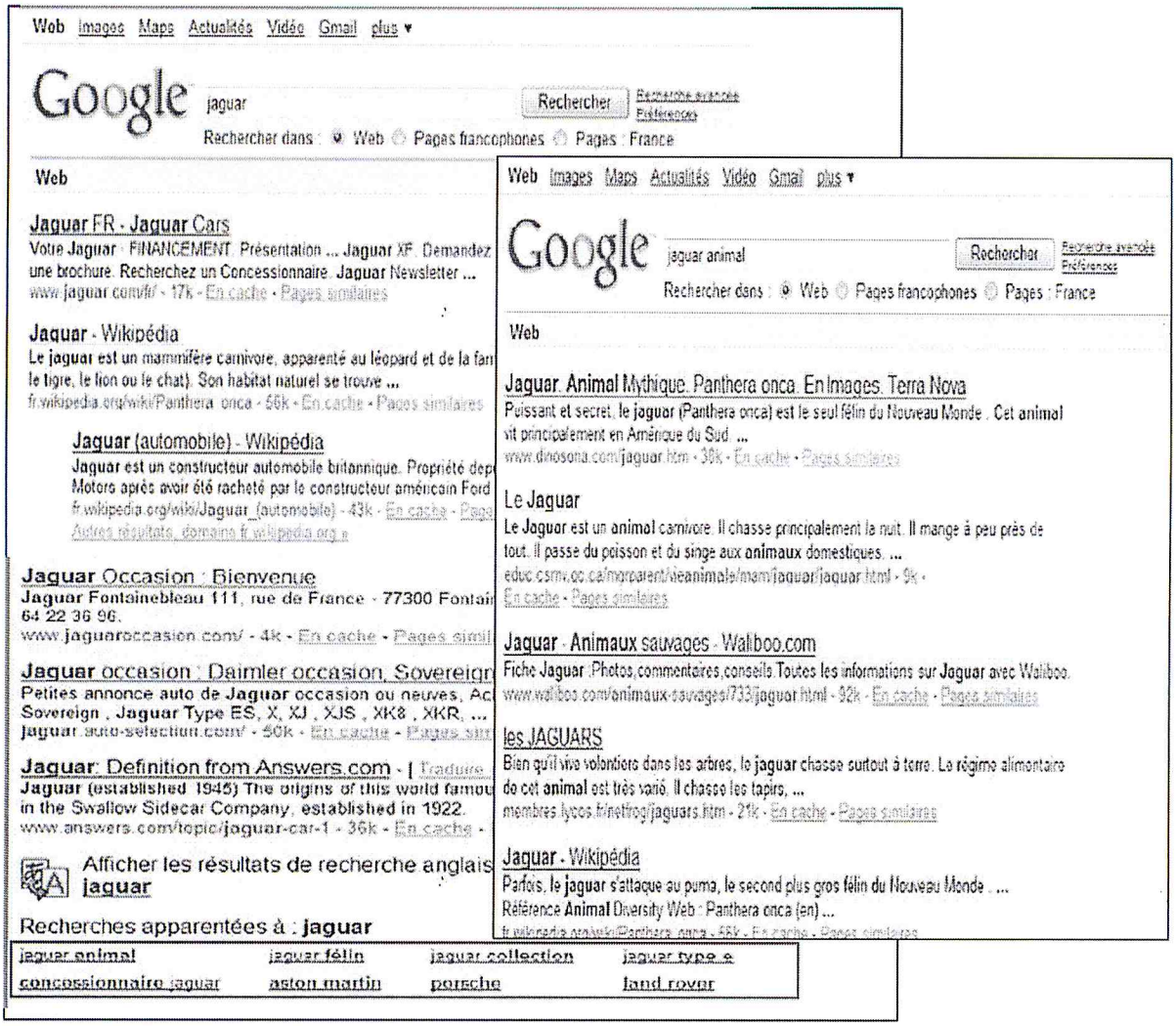

Figure 2: Désambiguïsation du sens des mots de la requête jaguar sur Google [2]

1.4.3. Regroupement thématique des résultats de recherche :

Face à la croissance du web et les difficultés rencontrées par les moteurs de recherche classiques pour satisfaire les besoins en information de l'utilisateur, les techniques de clustering/regroupement des résultats de recherche ont été développées pour une accessibilité et navigation plus simple. Les techniques de clustering sont basées sur le regroupement thématique (clustering) des résultats de recherche dans des catégories ou clusters à la place d'une liste de résultats paginés. Ces techniques sont basées sur le fait qu'un document qui est pertinent à une requête, a probablement une similarité avec d'autres documents qui sont peut être aussi pertinents. Ce regroupement permet de mettre les documents similaires ensemble et avoir une idée assez générale et globale des résultats retournés et ensuite une accessibilité et une navigation plus simple. Dans le même sens, plusieurs ontologies de domaines spécifiques ont été conçues dans le but de faire asseoir une recherche conceptuelle permettant de simplifier la navigation à travers les catégories sémantiques de la hiérarchie utilisée [2].

1.5. Recherche d'information sur le web :

1.5.1. Les outils de recherche d'informations :

Il existe de nombreux outils de recherche d'information sur le Web, ces outils se spécialisent en fonction des services utilisés et du type d'information qu'ils recensent. Il convient en effet de distinguer différents types d'outils de recherche sur l'Internet.
Un premier critère de classification des outils de RI repose sur le mode de recherche proposé. Il distingue entre les outils par navigation arborescente (comme les annuaires) ou hypertexte (comme les listes de signets), et les outils par requête (comme les moteurs, fondés sur l'utilisation de mots-clés). Cette distinction n'est plus pertinente aujourd'hui, tant l'imbrication est forte entre les mêmes outils.
Un deuxième critère reste toujours valable, en dépit des apparences : celui du mode d'indexation des ressources. Selon ce critère, on distingue les annuaires thématiques, qui procèdent à un référencement des sites Web et les moteurs de recherche, qui fonctionnent par collecte et indexation automatisées des pages Web (et non des sites). Cette distinction, 'historique', est moins nette aujourd'hui, à cause de l'imbrication des annuaires et des moteurs, *Google* utilise l'annuaire de l'Open Directory, *Yahoo* a son propre moteur, etc. [3]
On distingue trois catégories d'outils pour la recherche d'information sur le web: les moteurs de recherche, les annuaires et les méta-moteurs. Cette distinction qui repose également sur le mode d'indexation reste essentielle, car elle induit des usages et des technologies très différentes [3].

1.5.1.1. Les moteurs de recherche :

D'après [3], un moteur de recherche est une application permettant de retrouver des ressources (pages web, images, vidéo, fichiers, etc.) associées à des mots quelconques. Certains sites Web offrent un moteur de recherche comme principale fonctionnalité : on appelle alors moteur de recherche le site lui-même (*Google Video* par exemple est un moteur de recherche vidéo).
Ces outils de recherche sur le web sont constitué de « ***robots*** », encore appelés *bots, spiders, crawlers* ou *agents* qui parcourent les sites à intervalles réguliers et de façon automatique (sans intervention humaine, ce qui les distingue des annuaires) pour découvrir de nouvelles adresses (URL). Ils suivent les liens hypertextes (qui relient les pages les unes aux autres) rencontrés sur chaque page atteinte. Chaque page identifiée est alors indexée dans une base de données, accessible ensuite par les internautes à partir de mots-clés.
Les moteurs de recherche ne s'appliquent pas qu'à Internet : certains moteurs sont des logiciels installés sur un ordinateur personnel. Ce sont des moteurs dits ***desktop*** qui combinent la recherche parmi les fichiers stockés sur le PC et la recherche parmi les sites Web, on peut citer par exemple *Exalead Desktop*, *Google Desktop* et *Copernic Desktop Search*, etc.
Des modules complémentaires sont souvent utilisés en association avec les trois briques de bases du moteur de recherche. Les plus connus sont les suivants :

1. Le correcteur orthographique : il permet de corriger les erreurs introduites dans les mots de la requête et s'assurer que la pertinence d'un mot sera bien prise en compte sous sa forme canonique.
2. Le lemmatiseur : il permet de réduire les mots recherchés à leur lemme et ainsi d'étendre leur portée de recherche.
3. L'anti dictionnaire : utilisé pour supprimer à la fois dans l'index et dans les requêtes tous les mots "vides" (tels que "de", "le", "la") qui sont non discriminants et perturbent le score de recherche en introduisant du bruit.

1.5.1.2. Les annuaires :

L'annuaire (ou *directory* en anglais) est une liste de liens subdivisés en catégories suivant une structure en arbre, accompagnée d'une brève description. Les annuaires sont donc des outils basés sur le recensement humain de l'information. Ils signalent des sites et des ressources de l'Internet comme un catalogue de bibliothèque signale des livres ou bien encore comme les pages jaunes signalent des entreprises. On distingue dans ce contexte deux catégories d'annuaires [5].

A) Les annuaires commerciaux (Tableaux)

Ils se financent grâce à la publicité. Ils ont en principe une couverture dite "générale" (ils couvrent toutes les disciplines). Ils peuvent concerner le monde ou une zone régionale, nous citons parmi eux :

- Annuaires généralistes internationaux : le plus connu est sans doute 'Yahoo Directory', mais il existe aussi 'DMIZ' de l'Open Directory Project et l'annuaire de 'Lycos'.
- Annuaires régionaux commerciaux : ce sont les annuaires qui recensent des sites en fonction de leur langue. Dans le cas des annuaires francophones nous citons la version française de 'Yahoo Directory' ou encore l'annuaire 'Francité'.
- Les annuaires qui recensent d'autre pays ou parties du monde: comme l'annuaire 'Wohaa' pour l'Afrique et l'annuaire russe 'Yandex'.

B) Les annuaires non commerciaux

Sont des annuaires élaborés par des individus de façon bénévole ou bien par des institutions. Ils sont soit généraux soit spécialisés. Leur préoccupation consiste toujours à identifier les ressources et les sites en considérant leur qualité :

- Annuaires à couverture généraliste: comme le 'Vlib' (Virtual Library) et l'annuaire 'Resource Discovery network' (RDN).
- Annuaires à couverture thématique ou spécialisée : comme le répertoire en sciences humaines 'Voice of the Shuttle' et le répertoire de ressources juridiques 'Findlaw'.

1.5.1.3. Les méta-moteurs :

Ils sont de création plus récente. Ils constituent en fait la première génération des agents dits "intelligents". Ils permettent d'interroger en une seule fois différents outils de recherche , qu'ils soient de type annuaire ou de type moteur, afin de fournir une réponse plus exhaustive. Deux catégories de méta-moteurs: ceux en ligne et ceux consistant en un "logiciel client" à installer sur son ordinateur (le plus connu: COPERNIC). Le principe de fonctionnement des méta-moteurs est différent, Certains indexent l'information contenue dans différents annuaires et moteurs, d'autres les interrogent simultanément de façon dynamique. Certains de ces méta-moteurs retraitent plus ou moins les réponses (tri). Ils permettent ainsi de rechercher de façon plus large sur le Web. Toutefois, cela peut également générer du "bruit" (réponses non pertinentes). La parade mise en œuvre par certains méta-moteurs consiste à limiter le nombre de réponses de chaque outil interrogé (ce qui est indispensable et permet ainsi d'obtenir les réponses en principe les plus pertinentes) [3].

1.5.2. Architecture des moteurs de recherche :

Les architectures utilisées dans les systèmes de recherche d'information ont beaucoup évolué au cours de la dernière décennie. Cette évolution a été rendue nécessaire et même obligatoire car le nombre d'internautes ne cesse de progresser. De plus, les moteurs de recherche sont devenus le point de départ de beaucoup de sessions de navigation sur Internet et il a donc fallu trouver des solutions pour répondre en des temps raisonnables aux innombrables requêtes formulées à chaque instant.

1.5.2.1. Architecture générale des premiers moteurs de recherche :

L'architecture originale utilisée par Altavista représente la première catégorie de systèmes. Il s'agit d'une architecture très simple qui se divise en deux parties distinctes. On retrouve d'une part un crawler et d'autre part l'interface d'interrogation du moteur de recherche et le système d'analyse des requêtes proposés par les utilisateurs du système. Le crawler peut être considéré comme un robot chargé de rapatrier tous les documents Web contenus sur Internet dans un index centralisé en suivant les liens hypertextes rencontrés dans les pages analysées [10].
Le cœur du système repose sur un index inversé permettant d'associer des mots à un ou plusieurs documents. La demande de l'utilisateur est traitée en interrogeant l'index inversé pour connaître les documents dans lesquels apparaissent le plus souvent les mots de la requête [3].

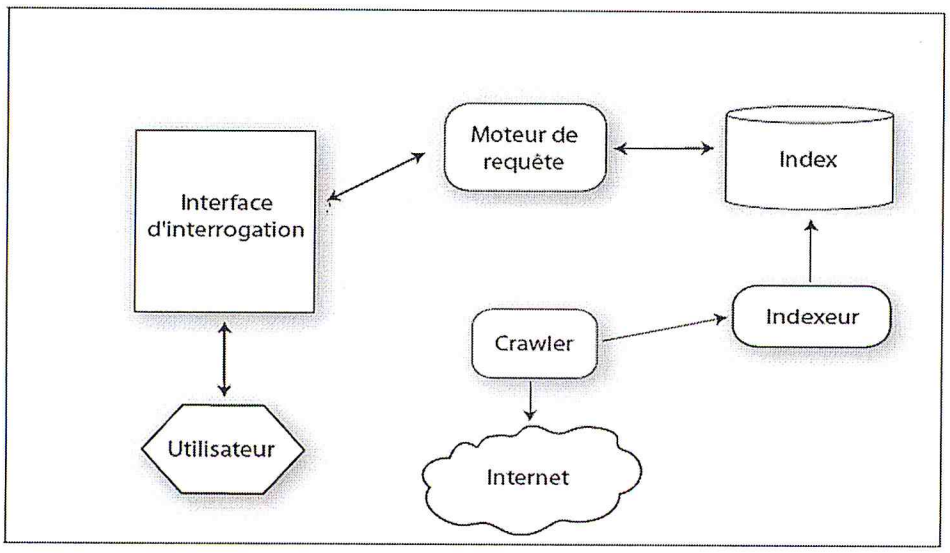

Figure 3: Architecture originale du moteur de recherche Altavista [10]

1.5.2.2. Vers un modèle distribué et adaptatif :

Des variantes de l'architecture précédente, basées sur le modèle indexeur-crawler, ont été imaginées afin de gommer les défauts inhérents à sa conception. L'une d'entre elle, appelée Harvest s'est révélée très innovante en matière de distribution des ressources.
Cette architecture a été utilisée par de nombreux organismes comme la NASA, l'Académie des Sciences Nationale des Etats-Unis d'Amérique.
Cette architecture se développe autour de deux composantes principales le récolteur et le broker. Chacun de ces éléments a un rôle particulier à jouer dans la chaîne de traitement du moteur de recherche. Le récolteur est chargé de collecter et d'extraire périodiquement

des informations d'indexation - textes, images - depuis plusieurs sites Web. Le broker, quant à lui, fournit le mécanisme d'indexation et l'interface d'interrogation sur les données amassées par le récolteur. On retrouve ici, le mécanisme indexeur-crawler identifié dans la section précédente. Cependant, plusieurs brokers et plusieurs récolteurs peuvent communiquer ensemble, chacun se spécialisant dans un domaine précis. Lorsqu'une requête est émise sur un broker dont le domaine traité ne correspond pas à ses capacités, celui-ci transmet la requête à une autre entité capable de la gérer.
C'est un système totalement adaptatif dans lequel il est possible de configurer les brokers et les récolteurs de manière à répartir le besoin en ressources sur un ou plusieurs domaines particuliers. Un système de réplication permet de plus de garantir une qualité de service relativement fiable [10].

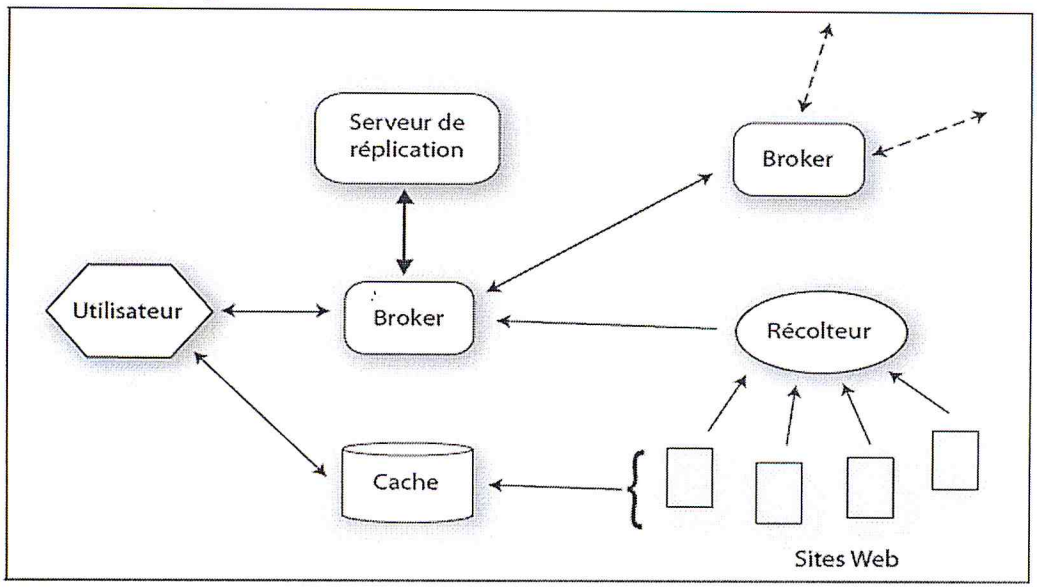

Figure 4: Architecture du système Harvest [10]

1.5.2.3. Architecture moderne d'un moteur de recherche :

L'architecture du moteur de recherche Google est certainement une des plus efficaces actuellement. Elle ne repose pas sur un système monolithique mais sur un grand nombre de machines classiques coopérant ensemble. Ce système peut se décomposer en plusieurs parties comprenant :
- Un sous-système d'exploration d'Internet
- Un indexeur
- Un analyseur de la topologie d'Internet formée par les liens hypertextes : et un sous-système de présentation et d'exécution de requêtes.
- Un serveur d'URL garde la mémoire des liens des pages à visiter. Des robots chargés d'explorer le Web récupèrent ces liens afin de télécharger les documents correspondant et les stocker dans une base de données recensant la totalité des pages indexées. Cette opération est réalisée continuellement et alimente et met à jour en permanence la base de documents du moteur. Périodiquement, cette base est analysée pour réaliser un index inversé reliant des termes aux documents les contenant. D'autres informations sur les termes sont extraites comme leur position dans le document, la taille de la police utilisée ou sa fonte [3]. Cet index inversé est distribué sur une multitude d'ordinateurs désignés par le terme barrel [10].

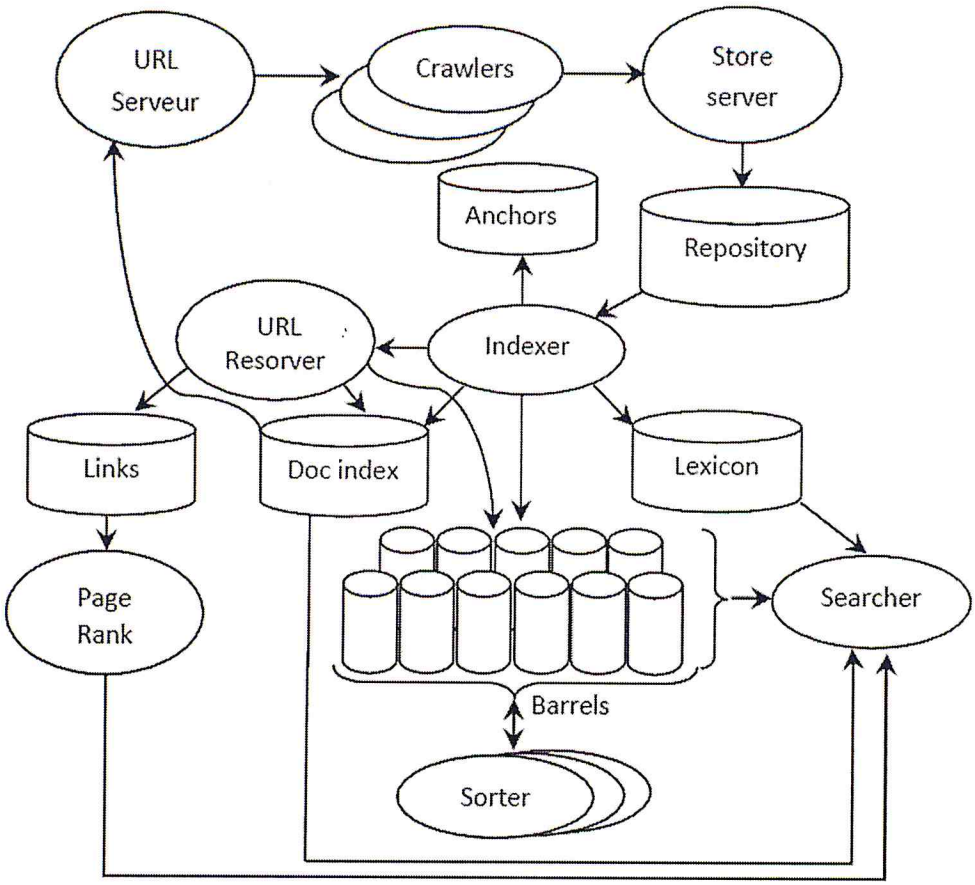

Figure 5:Architecture du moteur de recherche Google [3]

Cette analyse permet également d'extraire tous les liens hypertextes des documents rencontrés afin d'alimenter le serveur d'URL. Cette base de liens est utilisée afin de calculer le PageRank permettant de trier les documents de l'index par pertinence décroissante [3].

1.6. Conclusion :

Nous avons présenté au cours de ce chapitre les concepts de base de la RI classique, les systèmes de recherche d'information ainsi que l'évolution de la RI classique à la RI adaptative et on a terminé par les outils de recherche sur le web. Nous avons cité les techniques développées en RI adaptative, notamment la reformulation des requêtes, la désambiguïsation du sens des mots des requêtes et le regroupement thématique des résultats. Compte tenu des limitations de la RI adaptative, les travaux sont orientées vers la RI contextuelle dont le but est de mieux répondre aux besoins en informations de l'utilisateur selon une approche de RI non généraliste mais plutôt orienté contexte.

Chapitre 2

RI contextuelle

2.1. Introduction :

Compte tenu des limitations de la RI adaptative, les approches en RI se sont orientées vers une nouvelle génération de moteurs de recherche basés sur l'accès contextuel à l'information. L'objectif de la RI contextuelle est de mieux répondre aux besoins en information de l'utilisateur tout en intégrant le contexte de recherche dans la chaine d'accès à l'information [2]. Donc la RI contextuelle est une activité qui fait intervenir en grande partie l'utilisateur, assimilé en pratique à un ensemble d'éléments qui le décrive [13]. Parmi les éléments contextuels les plus importants traités dans la littérature, nous citons les centres d'intérêts de l'utilisateur connu par le contexte cognitif, sa tâche de recherche, ses préférences de recherche liées à la qualité de l'information retournée par le système (tel que la fraîcheur de l'information, la crédibilité de la source d'information, etc.), des préférences liées à la localisation géographique, au volume et au mode de présentation des résultats.

Ce chapitre traite la RI contextuelle. Nous essayons de faire la lumière sur la notion de contexte dans le cadre de la RI. Nous abordons les différentes possibilités de son utilisation en RI. Nous décrivons ensuite l'architecture générale d'un système d'accès contextuel à l'information. Ensuite nous parlons sur les différentes approches d'exploitation des dimensions du contexte dans la RI contextuelle, nous essayons de faire la lumière sur la RI personnalisée qui utilise le contexte personnel, ensuite nous passons àla RI contextuelle guidée par le contexte social, et nous terminons par la RI contextuelle guidée par le contexte mobile et le chapitre sera terminé par une conclusion à la fin.

2.2. Contexte et recherche contextuelle d'information :

La RI se fait selon un processus d'interaction homme-machine (H-M) où plusieurs facteurs interviennent et influencent la perception de pertinence del'information du côté utilisateur. Celui-ci apporte son jugement sur les documents renvoyés par le système selon des critères liés au contexte dans lequel la recherche est effectuée.
Il apparaît très vite que la notion du contexte est assez difficile à définir [3], et elle demeure floue à ce jour.

2.2.1. Définition du contexte en RI :

Le contexte n'est pas un concept nouveau en informatique : dès les années soixante, systèmes d'exploitation, théorie des langages et intelligence artificielle exploitent déjà cette notion. Avec l'émergence des systèmes de recherche d'information, le terme est redécouvert et placé au cœur des débats sans pour autant faire l'objet d'une définition consensuelle claire et définitive [3].

Les premières définitions de la notion de contexte en RI remontent aux travaux de Ingerwersen et de Saracevic [1] qui ont placé le contexteen amont de l'interaction utilisateur-SRI. Le contexte y est défini comme l'ensembledes facteurs cognitifs et sociaux ainsi que les buts et intentions de l'utilisateurau cours d'une session de recherche.
Par la suite, plusieurs définitions du contexte sont alors proposées dans la littérature, elles diffèrent essentiellement par ses éléments constitutifs et sa portée à court ou à long terme [1].
Le contexte à court terme inclut des éléments contextuels qui changent d'une recherche à une autre, tels que la localisation géographique de l'utilisateur, la nature de la tâche de rechercheou le type de besoin, etc. Le contexte à long terme inclut des éléments contextuels et des préférences de recherche persistants et évolutifs en mêmetemps, tels que les centres d'intérêts et les préférences liées àla qualité del'information [2].

Les premières approches en RI contextuelle se focalisent sur le contexte del'utilisateur représenté par son profil. Ce contexte inclut les centres d'intérêts,buts et connaissances de l'utilisateur dégagés au cours de ses sessions de recherche. Il a été démontré que les centres d'intérêts de l'utilisateur représententl'élément contextuel le plus important qui permet de résoudre l'ambiguïté de recherche dans un système de recherche d'information textuel [2].
D'autres travaux précisent que le contexte couvre des aspects largestels que l'environnement cognitif, social et professionnel dans lesquels s'inscrivent des situations liées à des facteurs tels que le lieu, le temps et l'applicationen cours, ou alors selon les familles de facteurs caractéristiques tels que leniveau environnement, le niveau utilisateur et le niveau interaction [1].Ainsi certains éléments du contexte peuvent être difficiles à cerner car nous les utilisons inconsciemment, d'autres se trouvent hors d'atteinte des périphériques d'entrée des machines et donc difficiles à mettre en œuvre dans des systèmes de recherche d'information [3].

2.2.2. Taxonomie du contexte :

Pour mieux comprendre les facteurs contextuels qui sont nécessaires en vued'une meilleure application du contexte dans les systèmes de RI, diverses taxonomies du contexte ont été proposées dans la littérature [1]. Ces taxonomies sont basées sur le concept du contexte multidimensionnel [2], faisant intervenir des dimensions cognitives dépendantes de l'utilisateur, spatio-temporelles dépendantes de l'environnement de recherche et qualitatives dépendantes dusupport de l'information.

2.2.2.1. Taxonomie de Fuhr 2000 :

Les troisprincipales dimensions retenues pour le contexte sont le caractèresocial, l'application et le temps [2].
1. la dimension sociale définit soit l'appartenance possible de l'utilisateur à un groupe ou à une communauté soit la non appartenance ce qui souligneun aspect individuel.
2. la dimension applicationdéfinit le but de la tâche accomplie en tantqu'une application workflow, une recherche adhoc ou une résolution de problème.
3. la dimension temps permet de définir le contexte temporel du besoin ;trois aspects temporels du besoin peuvent être identifiés : temps passé (batch), intention à court terme ou intention à long terme. Le contexte à court terme (interactif) est associé à des besoins et à des préférences instantanées de l'utilisateur alors que le contexte à long terme (personnalisation) traduit des besoins et des préférences persistants de l'utilisateur tout au long de diverses sessions de recherche. Une session de recherchepeut être associée à une itération de recherche représentée par une requête soumise et ses résultats associés ou alors un ensemble d'itérations de rechercheappartenant au même besoin en information.

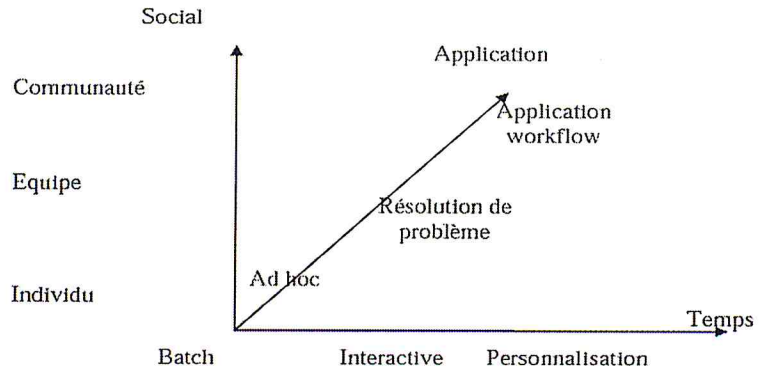

Figure 6:Dimensions du contexte multidimensionnel de Fuhr 2000 [2]

2.2.2.2. Taxonomie de Cool 2001 :

Une tentative de définition plus large des aspects du contexte est apparuedans le but de distinguer la notion de "contexte"de la notion de "situation" [2].Cette définition fait intervenir la tâche accomplie et les activités courantes de recherche ainsi que la dimension géo-spatiale de recherche dépendantdu lieu et du temps de recherche.Plus particulièrement, plusieurs aspects du contexte sont définis et sont liés à l'environnement cognitif, social et professionneldans lesquels s'inscrivent des situations liées à des facteurs tels quele lieu, le temps et l'application en cours. Dans une définition plus élaborée ducontexte, Cool et spink [2] définissent des niveaux contextuels considérés lesplus significatifs en RI afin de dissocier les entités intervenant dans le processusde recherche :

1. le premier niveau concerne l'environnement de recherche lié aux facteurscognitifs, sociaux ou professionnels qui influencent le comportement derecherche de l'utilisateur et sa perception de la pertinence.

2. le deuxième niveau concerne la RI liée aux connaissances de l'utilisateur,ses butset ses intentions de recherche.

3. le troisième niveau concerne l'interaction utilisateur-système et met en évidence l'impact des situations ou de l'environnement sur la rétroactionou les jugements de pertinence de l'utilisateur.

4. Le dernier niveau concerne le niveau de requête ou le niveau linguistiquedu contexte ; ce niveau explore la performance du SRI dans l'interprétation des requêtes des utilisateurs et leur habilité à les désambiguïser.

2.2.2.3. Taxonomie de Ingerwersen et al. 2005 :

Plus récemment, Ingwersen et Jarvelin ont développé une infrastructurecontextuelle cognitive dans le but d'étudier les dimensions du contexte quiont un impact sur le processus de RI [2]. L'infrastructure proposée consiste en 9classes/dimensions des contextes dépendantes l'une de l'autre. On cite :

1. La dimension de la tâche de travail naturelle dans une organisation qui concerne les caractéristiques d'intérêts liées à la tâche ;
2. La dimension de la tâche de recherche qui concerne les caractéristiques spécifiques à la tâche elle-même ;
3. La dimension de l'utilisateur qui concerne ses centres d'intérêts et sesconnaissances. Elle concerne aussi le type de besoin dans une tâche derecherche ainsi que sa perception vis à vis de cette tâche ;
4. Les caractéristiques de la collection (genre des documents, etc.) ;
5. Les caractéristiques du système telles que le principe de représentation desdocuments et des besoins ainsi que les méthodes d'appariement.

L'infrastructure propose aussi des caractéristiques d'interaction qui s'intègrent à la dimension d'accès à l'information. Elle définit également des variables temporelles concernant la durée des interactions (interactions à court terme, à long terme ou à base de session), le mode d'interaction (oral, iconique, pointage),etc.

2.2.2.4. Taxonomie de Tamine et al. 2009 :

Cette taxonomie comprend cinq dimensions principales :dispositif d'accès à l'information, contexte spatio-temporel, contexte de l'utilisateur,contexte de la tâche et contexte du document [11].

1. Contexte spatio-temporel : cette dimension comprend deux sousdimensions portant sur la localisation géographique et le temps.L'adaptation selon cette dimension concerne particulièrement les applicationsoù les informations ont une validité subordonnée au lieuet temps instanciés lors de l'activité de recherche d'information(routage, guide touristique etc.) [11].

2.Moyen d'accès à l'information : représente l'outil physique permettantd'effectuer unaccès direct à l'information tel que l'ordinateur,le téléphone portable, le PDA[1]etc. L'adaptation du processus de recherched'information aux caractéristiques de l'outil physique d'accèss'est imposé particulièrement pour des utilisateurs mobiles présentantdes contraintes d'ordre situationnel (requêtes courtes) etphysiques (ressources mémoires limitées, zone d'affichage des résultatsréduite) [11].

3. Contexte utilisateur : c'est la dimension principale abordée par lacommunauté. Cette dimension comprend deux sous-dimensions :contexte personnel et contexte social [11].

a. Contexte personnel : comprend à son tour les sous-dimensions suivantes :

– ***Contexte démographique*** : porte sur des facteurs de préférences personnelles tels que la langue et le sexe, exploitées de manière à personnaliser la recherchepour des besoins spécifiques.

– ***Contexte psychologique*** : l'anxiété et la frustrationsont desexemples de facteurs ayant un impact sur le comportementde l'utilisateur notamment son jugement de pertinence.

– ***Contexte cognitif*** : c'est laplus importante. Elle porte particulièrement sur l'expertise et les centres d'intérêtà court termeou centres d'intérêt à long terme de l'utilisateur.

b. Contexte social: cette dimension met l'accent sur la communautéà laquelle appartient l'utilisateur telle que les amis, les voisinset les collègues. L'adaptation du processus de recherche d'informationconsiste essentiellement à considérer les préférenceset les profils partagés par la communauté de l'utilisateurplutôt que ses préférences et profil personnels.

4. Tâche :cette dimension porte sur l'intention de l'utilisateurinduite par l'expression de sa requête [11]. Il existe trois(3) types de tâche [2] :

– **La tâche de recherche informationnelle :**qui s'inscrit dans le cadre de la recherche documentaire classique[11]. L'intention de l'utilisateur est de trouver del'information disponible sur le web dans une forme statique et dans plusieurs pages.Aucune interaction n'est prévue que la lecture. On donne comme exemple à ce typede requêtes celles qui demandent des informations dans les domaines de science,médecine, histoire, etc.[2].

– **La tâche de recherche navigationnelle** : dont l'intention est d'accéder à des sites d'accueil[11]. L'intention de l'utilisateur est de rechercher un sited'accueil qu'il a dans son esprit ou qu'il a déjà visité dans le passé ou bien ensupposant qu'un tel site existe.La plupart des requêtes qui contiennent des noms decompagnies, universités ou des organisations sont considérées navigationnelles[2].

– **La tâche de recherche transactionnelle** : dont l'intention est d'accéder à des services en ligne[11]. Le but général est d'accéder à un site où uneinteraction ultérieure va suivre. Cette interaction constitue la transaction définissantce type de requête. Les catégories principales de ce type de requête sont achats,trouver des services via le web, télécharger plusieurs types de fichiers (images,chansons, etc.), accéder à des bases de données (pages jaunes), trouver des serveurs(jeux), etc. [2]

[1]Personal Digital Assistant

5. Contexte du document (contexte de l'information) : Le contexte du document est issu du principe de la poly-représentation développé par Ingwersen. Le principe de la poly représentation est basé sur l'hypothèse que les documents peuvent être classifiés dans un espace à variablesmultiples (structure du document, genre du document, style du contenu, structure des hyperliens …) et que cette classification augmente la probabilité que ces documents soient utilisés dans certaines situations de recherche/ contexte derecherche. Trois sous-dimensions sont identifiées sous le contexte du document [2]:
– La première dimension concerne la représentation du document tel que les éléments structurels, la forme, les citations et les métadonnées.
– La deuxième dimension concerne les caractéristiques de la source des données, telles que sa crédibilité, sa fiabilité, etc.
– La troisième dimension concerne la qualité de l'information. Vu que la mesurede pertinence de l'information ne dépend pas seulement du contenu des données mais aussi de sa qualité, la notion de pertinence est étendue à une notion plus générale liée à la qualité de l'information retournée.
En général, la qualité de l'information concerne sa fraîcheur, sa précision, sa cohérence, sa complétude, sa sécurité, etc.

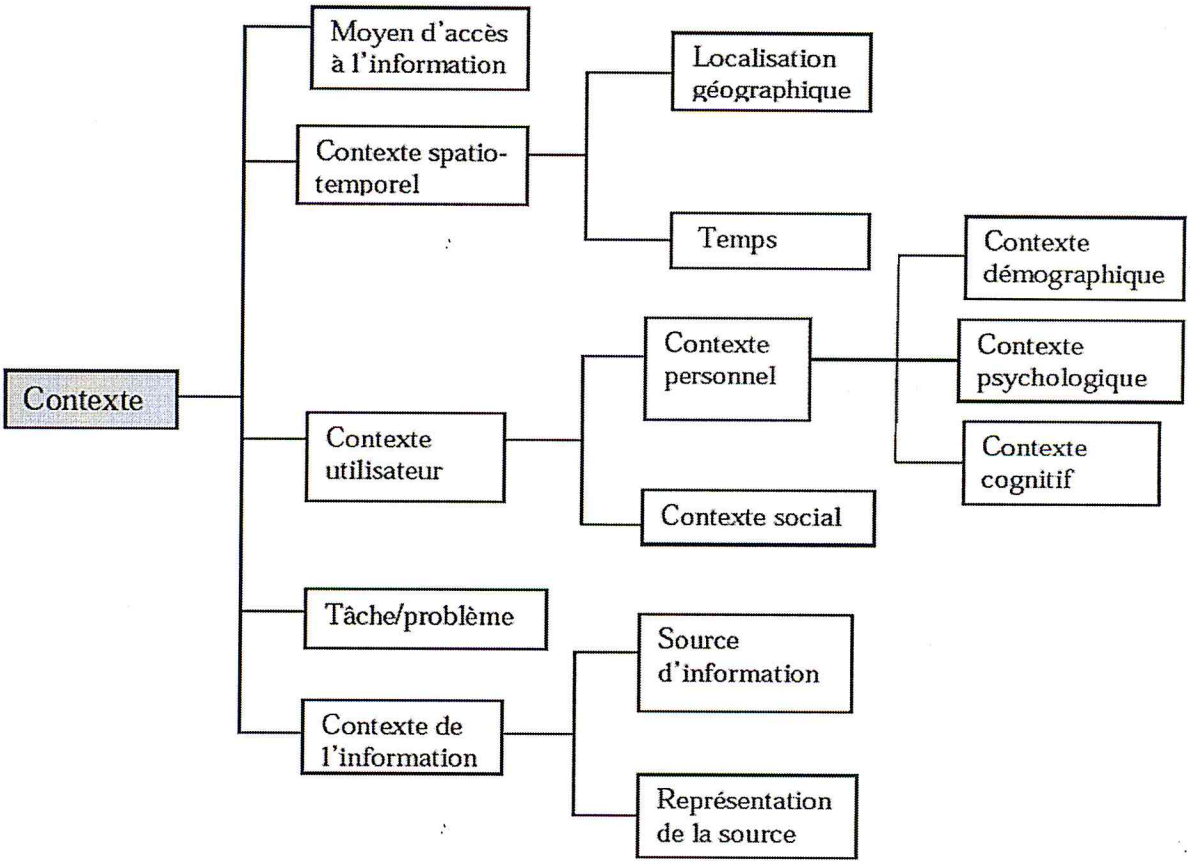

Figure 7:Le contexte : une notion multidimensionnelle [11]

2.3. Utilisation du contexte en recherche d'information :

En recherche d'information, le contexte peut être utilisé à trois stades différents selon l'avancement du processus de recherche lui même. Il peut donc être considéré au début duprocessus de recherche, au cours du processus de recherche, ou encore à la fin du processusde recherche [3].

2.3.1. Au début du processus de recherche :

Le contexte peut être utilisé dans une étape de pré-recherche pour résoudre le problème de l'ambiguïté des termes dans la requête et améliorer ainsi la qualité des résultats retournés par le système [3]. On peut aider l'utilisateur dans la formulation de sa requête en lui demandant de préciser, selon le contexte de la recherche en cours, le sens d'un terme ambigu en utilisant un thesaurus ou une ontologie. Une autre façon plus simple d'utiliser le contexte dans une phase de pré-recherche est d'utiliser le contexte spatio-temporel par exemple, un événement peut avoir lieu à 9h57, à 10h environ ou dans la matinée. Dans ce cas, le contexte peut servir pour le choix de la représentation appropriée.

2.3.2. Pendant le processus de recherche :

Le contexte peut également être considéré au niveau des interactions avec le système. En effet, dans un processus de recherche d'information, c'est l'interaction qui rend possible l'exploitation réelle de l'ensemble des résultats une fois affichés. L'intervention de l'utilisateur au cours du processus de reformulation est optionnelle. Dans le cas ou l'utilisateur précis des jugements par rapport aux résultats de la recherche, la procédure de reformulation est exécutée sur la base des documents jugés pertinents par l'utilisateur. Dans le cas contraire, une procédure permet d'estimer les préférences de l'utilisateur en terme de documents pertinents. Cette procédure analyse les documents restitués et applique un filtrage pour ne garder que ceux qui sont susceptibles d'êtres les plus pertinents et ça ce fait selon le contexte utilisateur. La procédure de reformulation automatique est alors exécutée sur la base de ces documents. [14]

2.3.3. A la fin du processus de recherche :

Durant la phase finale d'une session de recherche, l'intervention de l'utilisateur est obligatoire afin de faire face au problème de changement des intérêts. En effet durant une même connexion, l'utilisateur peut changer à plusieurs reprises son contexte de recherche. A chaque changement de contexte de recherche, l'utilisateur doit l'indiquer au système en annulant la recherche en cours, ou en sauvegardant le profil associé. Ainsi, le profil d'interrogation employé est modifié dès que l'utilisateur change de direction de recherche. L'intervention de l'utilisateur durant cette étape a pour but de valider le résultat de recherche qu'il vient d'obtenir. L'utilisateur doit préciser si le profil obtenu à la fin d'une recherche permet de retrouver l'information qui correspond à son besoin. Si c'est le cas le profil pourra être sauvegardé, afin d'être utilisé lors des prochaines interrogations. Sinon le profil ne pourra pas être sauvegardé. [14]

2.4. Système de RI contextuel :

Les systèmes de recherche sur le web traitent les requêtes isolées, les résultats pour une requête donnée sont indépendants de l'utilisateur, ou du contexte dans lequel l'utilisateur pose sa requête. On dit qu'un système de recherche d'informations est contextuel s'il utilise des données récupérées du contexte dans le but de délivrer l'information pertinente et appropriée. Ainsi la pertinence de l'information dépend de l'adéquation entre la requête et l'ensemble des éléments constituant le contexte qui sont perceptibles lors de la recherche.

2.4.1. Définition :

Un SRI est dit contextuel ou sensible au contexte (context-aware en anglais) s'il exploite les données du contexte de recherche pour sélectionner l'information pertinente en réponse à une requête utilisateur [3].La pertinence de l'information retournée à l'utilisateur est alors dépendante de son adéquation à la requête et de plus aux dimensions du contexte (définies ci-haut) qui sont perceptibles dans la situation de recherche en cours.

2.4.2. Architecture d'un système de RI contextuel :

On peut distinguer deux fonctionnalités fondamentales dans l'architecture de base d'un SRI contextuel : la modélisation du contexte et l'accès contextuel à l'information [1].

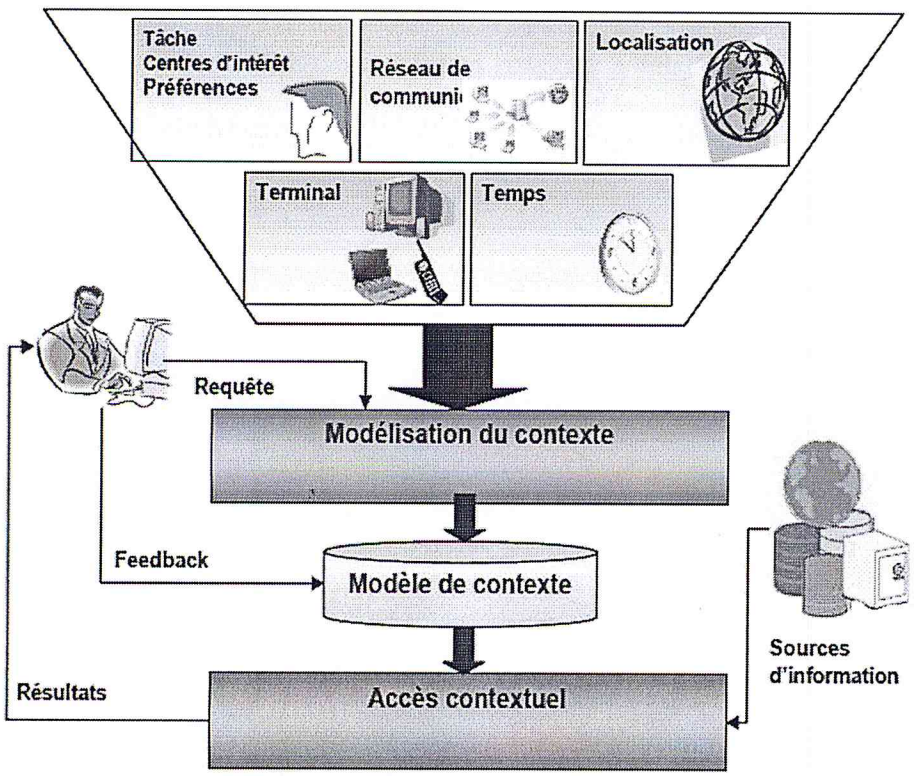

Figure 8:Architecture de base d'un SRI contextuel [11]

2.4.2.1. La modélisation du contexte :

La RI contextuelle s'appuie sur une source d'évidence additionnelle exprimée à travers le contexte qu'il convient alors de modéliser. La nature et la portée du modèle dépendent des dimensions du contexte considérées. Le contexte utilisateur étant la dimension la plus abordée, la modélisation du contexte est alors qualifiée souvent de modélisation de l'utilisateur [3]. De manière générale, un modèle de contexte est défini par l'instanciation des éléments suivants :

1.les sources d'information :environnement (temps, température etc.),collection de documents, historique des interactions … [11]Une revue de la littérature montre que les sources les plus utilisées sont les suivantes [1]:

– Le comportement de l'utilisateur perçu à l'aide d'indicateurs d'évaluationimplicite tels que l'historique des clicks, les données de navigationet le mouvement des yeux.

– Les pages et sites favoris.

– Des informations locales et contextuelles telles que les sources accédées comme les journaux, les Blog siteset les sites de e-commerce.

– Les premières pages ou résumés de pages web retournés par un moteur de recherche.

2.des stratégies de collecte de ces informations :on distingue principalement entre les stratégies implicites et stratégies explicites pour la collecte des données du contexte [3].

– L'acquisition explicite :repose principalement sur les techniques de feedback explicite largement utilisées dans la reformulation de requêtes par réinjection de pertinence.

Ces techniques d'acquisition explicites permettent une construction contrôlée du profil utilisateur. Cependant, elles présentent des limites à cause de l'effort supplémentaire imposé à l'utilisateur à spécifier explicitementses besoins [1].

– L'acquisition implicite : consiste à collecter à l'aide d'algorithmes d'acquisition implicite les données de l'utilisateur en observant ses interactions avec le système durant les activités de recherche. L'avantage de cette approche est qu'elle ne nécessite aucune implicationdirecte de l'utilisateur [1].

3. *des ressources de modélisation* : des ressources, généralement sémantiques(ontologies, dictionnaires, ...), sont parfois exploitées pourenrichir les données du modèle [11].

4. *des modèles de représentation et/ou évolution* :permettent de formaliser la représentation du contexte en qualité de structure unifiée (partie d'une ontologie, classe de vecteurs de termes, ensemble de concepts ...) ou d'un ensemble d'informations avec des structures différentes et spécifiques, puis de les faire évoluer en cours du temps [3].

Parmi les modèles de représentation proposés dans la littérature on peutdistinguer [1] :

– Les représentations basées sur l'historique de recherche :consistent en l'ensembledes requêtes et des pages web précédemment visitées ou cliquéesde l'utilisateur ou l'ensemble des requêtes et les résumés textuels de ses résultats associés accumulés au coursdes sessions de recherche de l'utilisateur.

– Les représentations ensemblistes : se basent sur un ensemble de mots clés (ou vecteurs de termes) pondérés représentés souvent selon le modèlevectoriel. Les paquets de termes représentent généralementles centres d'intérêt de l'utilisateur. Nous pouvons distinguer entre les représentations ensemblistes qui utilisent un vecteur de termes pondérésreprésentant un centre d'intérêt et celles qui utilisent desclasses de vecteurs de termes pondérés dont chacun représente un centred'intérêt.

– Les représentations connexionnistes : consistent non seulement à extraire des termes à partir des documents pertinents de l'utilisateur, mais à intégrer ces termes dans un réseau de nœuds

pondérés. Cette représentation permet de résoudre les failles de la représentation ensemblistepar la mise en place des relations de corrélation sémantiquesentre les mots du vocabulaire utilisé.
– Les représentations conceptuelles :se basent sur l'exploitation des ontologies des domaines ou des hiérarchies de concepts préalablement définies. L'approche de représentation conceptuelle consiste tout d'abord à spécifier les niveaux des concepts de l'ontologie à considérer et ensuite appliquer le procédé de déploiement des données dans des techniques de pondération de ces concepts. A la fin, le contexte utilisateur sera présenté par un réseau de nœuds conceptuels reliés entre euxen respectant la topologie des liens définis dans les hiérarchies ou les ontologies utilisées.

2.4.2.2. L'accès contextuel à l'information :

C'est le processus classique de RI projeté selon une dimension additionnelle liée au contexte de recherche. Principalement, son objectif est de sélectionner l'information pertinente à la requête adressée au SRI, en tenant compte la requête d'une part et le contexte de recherche en cours d'autre part [11]. Le contexte peut être exploité à différentes phases du processus de RI : dans la formulationde la requête, dans la fonction de pertinence, dans l'ordonnancement des résultats de recherche [1].

2.4.2.2.1.Reformulation de la requête :

Les éléments du contexte peuvent être utilisés pour reformuler une requête.
La reformulation de requête consiste à augmenter la requête avec des informationsdu contexte avant de lancer le processus d'appariement [1]. Des approches ont été développées dans ce contexte [1], en tenant exemple la reformulation de requête par intégration des termes représentant le contexte de l'utilisateur, et ça se fait par l'ajout des termes représentant le concept pertinent à la requête, et la suppression des termes représentant le concept non pertinent sélectionné par l'utilisateur. Une autre approche de reformulation de la requête [1] exploite un profil connexionniste, le processus de reformulation de la requête génère une nouvelle requête en appliquant toutes les écritures possibles définies par les arcs pondérés du profil et satisfaisants un critère du seuil de corrélation fixé.

2.4.2.2.2. Fonction d'appariement :

Le contexte peut également intervenir dans la définition de la fonction depertinence. Le calcul du score du document est alors une fonction qui assigneau document un score de pertinence en fonction non seulement de la requêtemais aussi du contexte utilisateur [1].

Des approches ont été développées aussi dans ce contexte [1], en tenant exemple la proposition d'une variante personnalisée de l'algorithmePageRank en l'occurrence PPV (Personalized PageRank Vector), son principe fondamental consiste à privilégier les pages reliées aux pages préférées de l'utilisateur ou les pages citées par ces dernières au cours du processus de calcul des scores de sélection [1]. Une autre approche qui intègre le contexte de l'utilisateur dans la fonction d'appariement du modèle bayésien de RI, l'approche proposée est basée sur l'utilisationdes diagrammes d'influence qui permettent de formaliser l'utilitédes décisions associées à la pertinence des documents compte tenu de la requêteet du contexte de l'utilisateur. Plus précisément, le score de pertinence d'undocument D est calculé à travers le diagramme d'influence, notéID (D, C, μ),où C est l'ensemble des centres d'intérêts $\{c_1, c_2, ...,$

$c_n\}$ modélisant le contextede l'utilisateur U et $\mu = \{\mu_1, \mu_2, ..., \mu_n\}$, où μ_k exprime l'utilité du document instancié D pour le centre d'intérêt c_k de l'utilisateur.

2.4.2.2.3. L'ordonnancement des résultats :

Cette phase peut également prendre en compte le contexte pour réordonner les résultats fournis par le processus de sélection. De ce fait, l'ordre final des documents à présenter à l'utilisateur est une combinaison du score/rang produit par le processus de sélection classique et celui fourni par la similitude avec lecontexte de l'utilisateur [1].

Des approches ont été développées aussi dans ce contexte [1], en tenant l'exemple de ré-ordonnancement des résultats de recherche par combinaison du score d'appariement originaldu document avec le score de similarité entre le document et les centres d'intérêtde l'utilisateur représentant son contexte, ce dernier est calculé en appliquant une mesure de similarité vectorielle basé sur le cosinus entre le document et le représentant du profil utilisateur.

2.5.L'exploitation des dimensions du contexte dans les modèles de recherche d'information contextuelle:

2.5.1.Accès contextuel à l'information guidé parle profil utilisateur:

La RI contextuelle guidée par le profil utilisateur (connue sous le nom de la RI personnalisée) est une branche de la RI contextuelle dont le contexte prendune dimension cognitive et défini par le profil de l'utilisateur.

2.5.1.1. La notion de profil utilisateur :

Le profil de l'utilisateur couvre des aspects larges tels que son environnementcognitif, social et professionnel qui déterminent ses intentions aucours d'une session de recherche. La plupart des travaux actuels en RIcontextuelle focalisent à juste titre, sur la représentation de l'aspect lié à cesintentions qualifiées de centres d'intérêts. Dans cette perspective, la modélisationdu profil de l'utilisateur a pour objectif fondamental de représenterpuis faire évoluer ses besoins en information à court et moyen terme. C'estune question qui pose la double difficulté de traduire les centres d'intérêt del'utilisateur d'une part et faire émerger leur diversité d'autre part.Le processus de définition du profil de l'utilisateur peut être caractérisé partrois phases. La première porte sur la représentation des unités d'informationreprésentant le profil. La deuxième phase est liée à l'instanciation dece modèle au cours d'une activité de recherche d'information. Enfin, la troisièmephase concerne l'évolution du profil au cours du temps. Chacune deces phases met en jeu des approches et techniques de représentation et/ou deconstruction résumées ci-après. [16]

2.5.1.2. Architecture fonctionnelle d'un système de RI personnalisé (SRIP) :[2]

Le but fondamental d'un SRI personnalisé est de satisfaire les besoins eninformation de l'utilisateur en intégrant son profil dans la chaine 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,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 etles centres d'intérêts de l'utilisateur ou un groupe d'utilisateurs. Un modèle de profilutilisateur 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 processusd'accès et retourner l'information pertinente en fonction de ce profil.

2.5.1.3. Les différentes approches de modélisation du profil utilisateur

Parmi les approches utilisées pour modéliser le profil utilisateur, on peut citer :

- **les vecteurs de mots clés** :[3] cette approche est basée sur un ensemble de vecteurs de mots clés pondérés représentés souvent selon le modèle vectoriel. Ce type de représentation est le premier conçu pour modéliser le profil utilisateur. Les mots clés traduisent les concepts auxquels l'utilisateur s'intéresse (les centres d'intérêts)et que le système doit prendre en compte pour délivrer les documents qui les contiennent.La construction d'un profil dans cette approche repose sur des techniques d'extraction des termes à partir des documents pertinents jugés implicitement ou explicitement par l'utilisateur.
- **les ontologies** : [4] les travaux actuels tendent à représenter le profil sous forme d'une ontologie deconcepts personnels en se basant sur les connaissances contenues dans les ontologies plutôt que de construire les profils d'utilisateur seulement à partir des documents collectés de son interaction.La représentation est essentiellement basée sur l'utilisation d'ontologies. Cette représentation peut également être assimilée à l'approchevectorielle du fait que les domaines sont souvent représentés comme des vecteurs de termes pondérés.Néanmoins, les termes de ces vecteurs sont regroupés pour former un domaine spécifique et non de simples mots-clés.De l'association des centres d'intérêts de l'utilisateur aux concepts des domaines de l'ontologie, on obtient un profil représenté sous forme d'une hiérarchie de concepts.
- **modélisation multidimensionnelle** :[4] Cette représentation a pour objectif de capturer toutes ces caractéristiques informationnelles de l'utilisateur. Donc plus de les centres d'intérêts de l'utilisateur (dimension centre d'intérêts), cette modélisation permet aussi de capturer autre dimensions et cela pour objectif de sécuriser les profils, ces dimensions sont les attributs démographiques del'utilisateur (identité, données personnelles), les attributs professionnels (employeur, adresse, type) et les attributs de comportement (trace de navigation).

2.5.1.4. Les différentes approches d'acquisition des données utilisateurs

On distingue deux approches d'acquisitions, explicite et implicite [6]:

- **l'acquisition explicite** : cette technique constitue une approche simple pour obtenir des informations sur l'utilisateur. On interroge directement l'utilisateur ou on lui demande par exemple de remplir des formulaires pour collecter les données personnelles et démographiques tels que sa date de naissance, son statut marital, son activité professionnelle et ses centres d'intérêt.
- **l'acquisition implicite** : cette technique consiste à collecter les informations décrivant l'utilisateur, en observant les dimensions et les membres fréquemment sollicités et en scrutant les caractéristiques de l'environnement à partir duquel il intervient (les

capacités et les limites du dispositif utilisé lors de ces interactions). Et ce, en se basant sur l'historique de ses interactions avec le système.

2.5.1.5. Les différentes approches d'exploitation du profil utilisateur

On distingue trois approches principales qui consistent à intégré le profil utilisateur dans le processus d'accès à l'information [5] :

1- reformulation des requêtes.
2- appariement documents-profils.
3- ré-ordonnancement des résultats de la recherche.

- **reformulation des requêtes** :le but fondamental de la reformulation de requêtes par utilisation de profil consiste à cibler la recherche des documents pertinents par augmentation de la requête par des termes issus du profil utilisateur dans le but de mieux répondre au besoin en information de l'utilisateur [3].
- **appariement documents-profils** :les modèles d'appariement personnalisés consistent à exploiter le profil utilisateur dans la fonction de calcul du score du document vis-à-vis une requête. La fonction classique de calcul du score du document se base sur la requête comme la seule ressource d'information qui représente l'utilisateur. Dans le cadre de la RI personnalisée, le calcul du score du document est une fonction qui assigne au document un score de pertinence en fonction non seulement de la requête mais aussi du profil utilisateur [3].
- **ré-ordonnancement des résultats de la recherche** :le principe du ré-ordonnancement est de modifier l'ordre de l'affichage des résultats au client. Il s'agit d'un post traitement qui étant donné les éléments retournés par une requête, essaie de trouver une manière d'échanger leurs emplacements en fonction des préférences de l'utilisateur. L'échange de l'ordre d'apparition des éléments des résultats est fait généralement en appliquant une fonction qui permet de calculer le nouveau rang de l'objet [2].

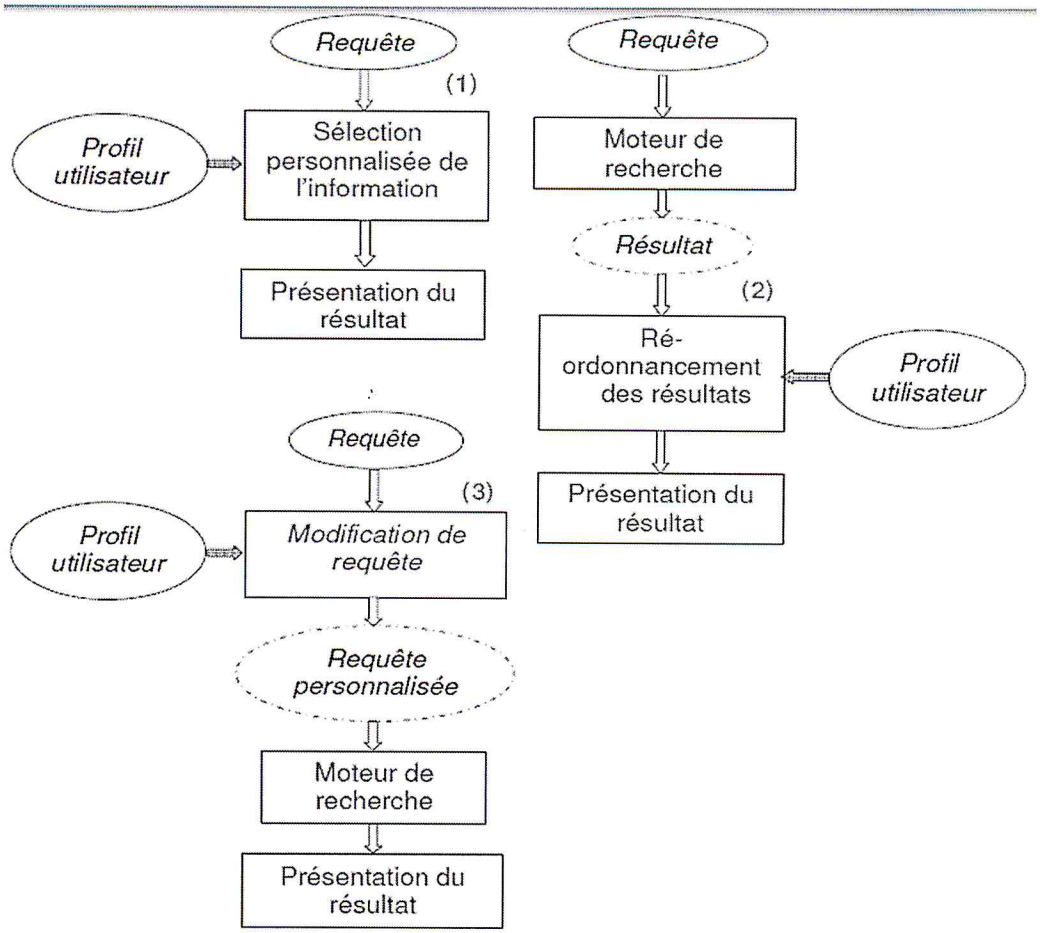

Figure 18:- phases d'intégration du profil utilisateur dans le SRI[4] -

(1) : Modèle d'appariement documents-profils
(2) : Ré-ordonnancement des résultats de la recherche
(3) : Reformulation des requêtes

2.5.2. Accès contextuel à l'information guidé par le contexte social :

La recherche sociale d'information (RSI) est fondée sur le fait que l'on ne peut pas séparer le producteur d'une information de son produit et que nous utilisons les gens pour trouver des contenus et le contenus pour trouver des gens [17].
La RSI est à l'intersection des concepts de la recherche d'information et des réseaux sociaux. Elle comporte toutes les techniques permettant à un utilisateur d'avoir accès aux informations susceptibles de répondre à son besoin informationnel, tout en exploitant les connaissances d'autres utilisateurs. Parmi ces techniques se trouvent l'indexation sociale, le *bookmarking* social, le filtrage collaboratif, l'analyse des réseaux sociaux et le partage de requête [18].

2.5.2.1. Système de recherche social d'information :

Un système de recherche sociale d'information (SRSI) comporte trois éléments : document, requête, individu. Ce qui domine les systèmes de RI traditionnels est la modélisation des relations entre les documents et les requêtes. Les SRSI, de leur côté intègrent en plus la modélisation des relations entre les individus, ce qui fait leur force. Dans

les SRSI, l'individu peut jouer le rôle de producteur de l'information ou celui d'utilisateur de l'information, et nous prendrons garde à bien faire cette distinction [18].

2.5.2.2. Quelque techniques dans le domaine de la RSI :

2.5.2.2.1. Le social bookmarking :[19]

Une parmi les techniques proposées dans le domaine de la recherche d'information sociale (contexte social), est d'utiliser le social bookmarking afin d'améliorer le processus de la recherche d'information.

a. Tag :

Le terme « tag » (en français « étiquette ») représente unmot-clé ou une expression associée ou assignée aux ressources. Il décrit ainsi l'objet et lui permetd'être retrouvé par navigation, par filtrage ou par recherche. L'activité des individus concernant l'attribution des métadonnées aux ressources s'appelle «tagging».
L'ensemble des tags d'un individu forme une collection qui s'appelle « personomy ». L'ensembledes personomies constitue la «folksonomy». Elle est constituée par l'effortcollectif des utilisateurs qui ont diverses connaissances et besoins quand ils interagissent avec unsystème de *social bookmarking*. De plus, le terme *folksonomy* est une combinaison de«folk» et «taxonomy», décrivant le phénomène de classification sociale, les ressources,les utilisateurs, les tags, ainsi que l'activité d'un individu. Cette activité consiste à créer uneassociation entre les utilisateurs et les ressources à travers des étiquettes.

La recherche d'information peut bénéficier de l'activité de *social bookmarking*. Les tags permettent aux individus d'organiser leurs ressources d'une façon constructive,ce qui leur apporte des bénéfices dans la recherche. Une technique récente appelée «tag cloud» est développée pour visualiser les tags avantageusement et facilement par les utilisateurs dans le processus de recherche d'information.

a.1.Tag Cloud : visualisation des tags les plus populaires :
Pour faciliter la recherche d'information dans les grandes bases documentaires, certains systèmesde *social bookmarking* fournissent une interface appelée *Tag Cloud*.
Le *Tag Cloud* est une liste des étiquettes les plus populaires,généralement affichée dans l'ordre alphabétique et visuellement pondérée par la taille de police.
Dans un *Tag Cloud*, quand un individu clique sur une étiquette, il obtient une liste des ressourcesqui ont été taguées avec cette étiquette, ainsi qu'une liste des étiquettes qui y sont reliées.

Figure 10:Tag Cloud du site de social bookmarking « Del.icio.us »

b. Bookmark :

De nos jours, la quantité de données disponibles sur le web devient tellement volumineuseque l'utilisateur se trouve face à une masse d'information difficile à appréhender. Cette massed'information induit souvent une surcharge cognitive chez l'utilisateur et rend difficile l'accès àl'information répondant à ses besoins. Par conséquent, la recherche d'une information spécifiquesur le Web peut être une vraie gageure.

Une solution pour alléger la surcharge d'information peut consister dans le développementpar chaque utilisateur d'un système personnel d'information qui représente un sous-ensemble cibléd'informations pertinentes pour lui. Le bookmark représente un outil simple pour construire, ces sous-ensembles d'information personnalisés où des pages Web identifiées par des URLs utilesou intéressantes pouvant être stockées afin de s'en servir pour une utilisation ultérieure. Les utilisateursgardent la trace des liens vers des pages en créant une liste des bookmarks : un espacepersonnel de stockage d'informations Web.

Lesbookmarks sont utilisés comme des « espaces d'information personnelle du web » pour aider lesgens à se rappeler et à récupérer des pages Web.

Les bookmarks réduisent la surcharge physique et cognitive de gestion des URLs, en facilitantle stockage, la gestion et l'interprétation des liens (les utilisateurs ne doivent pas taper leslongues adresses), en aidant la mémoire et en gardant l'historique.Les limites qui s'imposent dans ce contexte ont trait à la difficulté de partager les ressources sauvegardées dans les navigateurs internet (favoris pour IE et bookmarks pour Mozilla) avecla communauté et avoir accès aux bookmarks sur n'importe quel ordinateur. Dans la sectionsuivante nous présentons la notion *social bookmarking* qui répond à ces deux limites.

c. *Social bookmarking :*

Ce concept définit le moyen par l'intermédiaire duquel les internautes trouvent, accumulent,catégorisent et contrôlent l'information à partir des pages web.

Dans un système*social bookmarking*, les individus ont la possibilité d'étiqueter chaquelien qu'ils sauvegardent avec des mots-clés (tags).

Les sites de *social bookmarking* facilitent la navigation et l'accès aux informations en rendantplus rapide la recherche d'information sur le Web et favorisent la collaboration, la publicationet l'archivage des pages Web en facilitant la création d'un espace personnalisé de stockage d'information.Ainsi que, la publication et l'étiquetage des pages avec les auteurs et le partage despages en favorisant la collaboration en utilisant les documents sur l'internet.

Dans le cadre des systèmes de *social bookmarking* trois types de bookmarks sont proposés :

c.1. Bookmark public :

Ce sont des bookmarks que les personnes décident de rendre visibles à la communauté.

Une communauté rassemble les individus qui détiennent un compte sur un site de *socialbookmarking*, mais aussi les personnes qui ne sont pas membres.

c.2. Bookmark privé :
C'est un bookmark visible seulement pour la personne qui a posté saréférence.

c.3. Bookmark partiellement visible :
La personne qui a sauvegardé la ressource peut établir le degré de visibilité pour le bookmark (visibilité restreinte à un groupe d'utilisateurs, par exemple).

2.5.2.2.2. Le filtrage collaboratif [20]:

Le filtrage collaboratif est une autre technique qui s'inscrit dans la recherche sociale d'information. Il regroupe les méthodes pour le développement du système de recommandation en se fondant sur les préférences et les apports de la communauté d'utilisateurs du système. Le filtrage collaboratif permet de contourner certaines difficultés liées au système de filtrage par le contenu et au système de recherche d'information. Le principe est de filtrer le flot de documents entrant en fonction de l'opinion que d'autres utilisateurs de la communauté ont déjà portée sur les documents.

Dans les systèmes de filtrage collaboratif, les méthodes statistiques sont utilisées pour faire des prévisions basées sur des *intérêts des utilisateurs*. Ces prévisions sont à leur tour exploitées pour faire des propositions à un utilisateur individuel, en se fondant sur la corrélation entre son propre profil personnel et les profils d'autres utilisateurs qui présentent des intérêts et des goûts semblables. Pour construire leur profil, les utilisateurs *fournissent des évaluations* sur les objets informationnels.

Les évaluations peuvent prendre la forme de notes (scalaires) ou une forme binaire telle que d'accord / pas d'accord ou une forme unaire telle qu'un enregistrement ou une trace qui montre qu'un utilisateur a choisi un item ou a réalisé un achat (dans le domaine du commerce), les évaluations des utilisateurs sont comparées pour mesurer leurs similitudes. Des prévisions sont calculées sous forme de moyennes pondérées, à partir des évaluations d'autres utilisateurs ayant des goûts semblables ou complètement opposés.

Pour qu'un système de filtrage collaboratif soit efficace, deux conditions sont nécessaires : *ungrand nombre d'utilisateurs et une évaluation*, au moins, de chaque document du système. Des problèmes apparaissent alors pour les nouveaux documents. En effet, ils ne peuvent être diffusés que lorsqu'un minimum d'informations les concernant est collecté à partir de l'évaluation d'au moins l'un des utilisateurs. La limitation majeure donc, du système de filtrage collaboratif est liée au démarrage à froid. Les nouveaux utilisateurs commencent toujours avec un profil vide. Même s'il existe un profil de démarrage, le système a besoin d'une période d'apprentissage avant que le profil ne reflète correctement les préférences de l'utilisateur. En d'autres termes, le système ne peut pas filtrer des objets efficacement pour le compte d'un utilisateur pendant qu'il « apprend » sur lui. D'un autre côté, les personnes ayant des goûts peu fréquents risquent de ne pas recevoir de propositions.

Le premier système de filtrage collaboratif est le **système Tapestry** développé au centre de recherche Xerox Palo Alto (PARC). Il permet aux utilisateurs d'annoter les objets informationnels en texte libre ou de donner des appréciations dans le style « J'ai bien aimé » ou « Je déteste ». Ainsi les utilisateurs peuvent se recommander des documents les uns aux

autres. D'autres systèmes de filtrage collaboratif parmi les premières générations sont **Grouplens**, **PHOAKS** (People Helping One Another Know Stuff), **Siteseer** et **Fab**.

Parmi les systèmes commerciaux en ligne qui emploient la technique du filtrage collaboratif, nous trouvons :
- **Amazon**[2] qui est notamment connu pour la recommandation des livres ;
- **Amie Street**[3] qui permet aux utilisateurs de recommander des musiques. D'un autre côté, il leur permet de découvrir des musiques qui peuvent les intéresser en leurs fournissant des recommandations selon leurs préférences. Il permet également un système collectif de calcul de prix d'une musique ;
- **eBay**[4] qui est un site d'e-commerce de ventes aux enchères permettant aux utilisateurs de vendre et d'acheter des biens. Il recommande aux utilisateurs des itemsselon leur profil ;
- **Google News**[5] qui est un site de journaux en ligne, rassemble des articles provenantde plus de 4,500 sources et regroupe les informations similaires pour les afficher en fonctions des intérêts de chaque utilisateur ;

2.5.3. Accès contextuel à l'information guidé par le contexte mobile :

L'évolution de l'équipement et des usages ainsi que l'évolution technique des terminaux et des navigateurs et enfin l'évolution des forfaits proposés par les opérateurs, sont les trois facteurs majeurs qui ont permis le démarrage de l'Internet Mobile [1]. Au-delà des sites web mobiles, simples versions adaptées de sites web existants, des services exclusivement dédiés à une utilisation spécifique sur mobile sont aussi apparus et leur nombre ne cessent d'augmenter. Par conséquence, il est apparu une demande croissante pour des outils de recherche efficaces adaptés aux environnements mobiles. Le concept de système ou de moteur de recherche mobile est alors apparu, il est défini comme un logiciel conçu pour un appareil mobile pour fournir un service ou un portail, par lequel l'utilisateur peut soumettre une requête (habituellement par l'entrée d'un ensemble de mots clés) et obtenir la liste des résultats correspondants aux critères de recherche [1].

2.5.3.1. Définition de la RI mobile :

La RI mobile consiste en l'indexation et la recherche d'information textuelles et multimédia pour être retournées sur un dispositif mobile avec des connections sans fils [1].

2.5.3.2. Notion de contexte dans la RI mobile :

Les premières définitions du contexte dans l'informatique mobile défini le contexte par : « le contexte est défini par toute information qui peut être utilisée pour caractériser une entité. Une entité peut être une personne, un lieu, un objet, pouvant être considéré comme approprié dans l'interaction homme/application, incluant l'utilisateur et l'application eux mêmes. » [1]
Une autre définition qui défini le contexte mobile par : « Le contexte est un ensemble d'états et de paramètres qui soit détermine le comportement d'une application ou bien dans lequel un événement d'application se produit et est intéressant pour l'utilisateur.» [1]

[2] http://www.amazon.com
[3] http://amiestreet.com
[4] http://www.ebay.com
[5] http://news.google.com

Autre définitions qui définissent le contexte par la localisation, la proximité d'autres personnes, la température, le jour, etc. [1]

2.5.3.3. Construction du contexte mobile :

a. Le contexte temporel :le contexte temporel est le plus facile à obtenir en utilisant l'horloge intégrée au système qui est habituellement disponible dans tous les appareils mobileset fournit l'heure exacte.Le mode d'acquisition des sources d'information pour ce type de contexteest implicite et automatique. [1]

b. Le contexte spatial :pour l'acquisitionde la localisation des utilisateurs, un large éventail de différentestechnologies de positionnement sont utilisées. Ces technologies dépendentselon que l'application de RI fonctionne à l'intérieur ou à l'extérieur. Pour un positionnement extérieur, le système de positionnement mondial(Global Positioning System-GPS) ou son amélioration le GPS différentiel sont largement utilisés. Le GPS est un système de navigation par satellitesdisponible au niveau mondial qui permet aux périphériques compatibles de déterminer leurs positions, la vitesse et la direction du mouvement deleurs utilisateurs [1].Toutefois, dans des espaces d'intérieur le GPS manque de fiabilité etd'exactitude. Son signal de faible intensité est facilement bloqué par laplupart des bâtiments et en outre perturbé par les réflexions. Pour cette raison, différentes technologies sont apparues pour faire face au problème du positionnement à l'intérieur. Des systèmes qui exploitent latechnologie de positionnement par infrarouge pour couvrir les situationsd'utilisation à l'intérieur.Autre système exploite les identificateurs des antennes cellulaires (ou station de base) de téléphonie mobile pour identifier la localisation de l'utilisateur [1].
Le mode d'acquisition des sources d'information pour le contexte spatialest implicite. Cependant, il n'échappe pas à la connaissance de l'utilisateur qui peut autoriser ou bloquer l'accès au service de localisation [1].

2.5.3.4.Présentation d'une approche dans le domaine de la RI mobile : [1]

Plusieurs approches ont développées dans le domaine de la recherche d'information contextuelle mobile (contexte mobile) une parmi eux est caractérisée par :
-la combinaison de l'adaptation à la localisation, au temps et aux centres d'intérêt de l'utilisateur.
- l'exploitation à la fois de représentations brut es (récupérées des capteurs mobiles) et sémantiques (des concepts d'ontologies) du contexte.
- la construction de profils situationnels à base de l'historique de recherche annoté par le contexte spatio-temporel.

a. Approche générale :

L'utilisateur est engagé dans une situation.Lorsqu'il a besoin d'information, il soumet une requête à un moteur de recherche traditionnel, qui retourne une liste de documents. Pour adapter les résultats de recherche aux centres d'intérêt de l'utilisateur mobile, à sa localisation et au temps, un processus de contextualisation est mis en œuvre, il consiste à construire une représentation de ces éléments contextuels de l'utilisateur et des documents. Puis un processus d'appariement permet de définir une fonction d'appariement pour chaque type de contexte et de combiner les différents scores contextuels. Les documents seront réordonnés selon ce score

contextuel en plus de leur score initial. L'utilisateur sélectionne les résultats qu'il juge pertinents pour sa situation et l'historique est alors mis à jour.

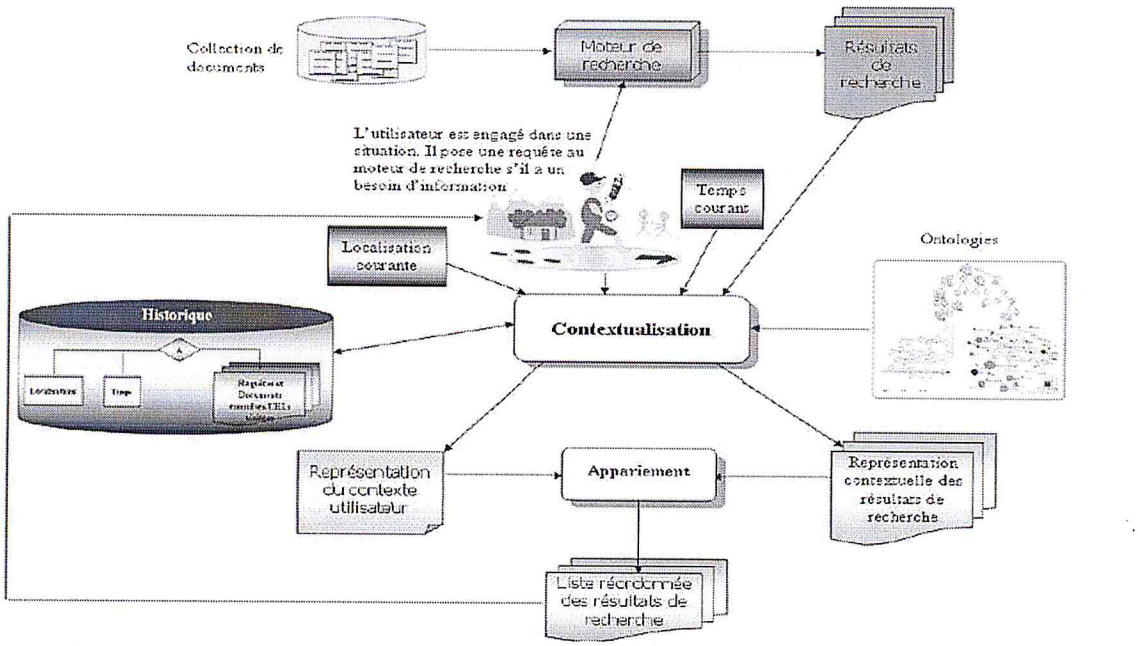

Figure 11: Schéma général de l'approche dans le domaine de la RI mobile.

b. Processus de contextualisation :

Il consiste à créer les représentations du contexte : centres d'intérêt de l'utilisateur, localisation et temps pour l'utilisateur et pour le document.

b.1. Les centres d'intérêt :

Le but est d'affiner la notion de centre d'intérêt vers des besoins situationnels de l'utilisateur. En effet, la mobilité induit des besoins informationnels typiquement conséquents de la situation dans laquelle se trouve l'utilisateur. Nous pensons qu'une partie des informations quiaident à clarifier une situation dont découle le besoin en information, sont les aspects contextuels de son moment d'émission, à savoir le temps et la localisation, exemple : "être la nuit à la maison", "l'été à la plage", etc. Cependant, il est clair qu'une situation ne se caractérise pas par des pures coordonnées géographiques et des points temporels. Par exemple la date n'est pas forcément l'attribut à utiliser pour identifier une situation, des attributs tels que le moment de la journée, le jour de la semaine, la saison … montrent plus de pertinence pour la caractérisation d'une situation (le même raisonnement s'applique à la localisation). C'est pour cela que nous proposons d'associer les informations du contexte récupérées par les capteurs du mobile à des concepts sémantiques (extraits d'ontologies temporelles et spatiales) pour pouvoir récupérer toutes les propriétés qui décrivent au mieux les facettes temporelles d'une date et d'une heure et les facettes spatiales des pures coordonnées géographiques.
Une situation sera déterminée par une classification sémantique d'une agrégation de points du contexte spatio-temporel. La construction des profils se fera ensuite sur la base de l'historique des recherches attachée aux situations identifiées. On se basera sur une représentation sémantique des profils par des concepts d'ontologie.

b.2. La localisation :

Pour pouvoir représenter la localisation de l'utilisateur au moment de la requête et caractériser sa situation (nom et type de place), en plus de réaliser une indexation spatiale des documents, un modèle pour la représentation dè localisations est nécessaire. L'information géographique peut être représentée à divers niveaux de granularité et sous diverses formes. Pour permettre une bonne représentation de l'information géographique et sa manipulation, la tendance est actuellement vers des approches sémantiques avec des ontologies spatiales. Nous proposons de se baser sur une base de données spatiale et un thesaurus spatial pour représenter et raisonner sur les données géographiques.

b.3. Le temps :

Pour définir les aspects temporels liés au temps de la requête et à la situation de l'utilisateur (matin, soir, *weekend*,...), mais aussi aux documents (heures et/ou dates de disponibilité), un modèle pour la représentation du temps est nécessaire. L'information temporelle est une information complexe, elle est continue et peut être représentée à divers niveaux de granularité. Pour permettre une bonne représentation de l'information temporelle et sa manipulation, la tendance est actuellement vers des approches sémantiques avec des ontologies temporelles.

c. Processus d'appariement :

Une fois que le contexte est représenté, il est exploité dans le processus de RIafin d'améliorer la pertinence de l'information retournée. Notre exploitation consisteà réordonner les résultats de recherche en tenant compte plusieurs scorescontextuels: un score de personnalisation, un score géographique et un scoretemporel, en plus du score initial d'un document. A chaque fois que l'utilisateursoumet une requête au moteur de recherche, sa localisation et son temps courantsont récupérés. Ils seront utilisés d'une part, pour identifier la situation courante envue de choisir le profil adéquat pour la personnalisation des résultats de recherche.Et d'autre part, pour calculer un score de pertinence géographique et/ou temporeldes documents retournés.

2.6. Conclusion :

Nous avons présenté dans ce chapitre la RI contextuelle. En RI contextuelle, le contexte dénote pratiquement tout élément qui peut affecter le jugement de la pertinence. L'importance du contexte s'explique par sa capacité à donner sens au phénomène observé. C'est effectivement en regard aux éléments environnants, aux conditions favorisant l'émergence de la requête et des documents que leur sens est révélé. Plusieurs études en sciences cognitives s'accordent dans ce sens : l'information doit être comprise dans son contexte et la prise en considération du contexte dans la RI est cruciale.

Chapitre 3

Reformulation de requêtes et Réinjection automatique de pertinence

3.1. Introduction :

La reformulation de requête est proposée comme une méthode élaborée pour la recherche d'information s'inscrivant dans la voie de conception des SRI adaptatifs aux besoins des utilisateurs. C'est un processus permettant de générer une requête plus adéquate à la recherche d'information dans l'environnement du SRI, que celle initialement formulée par l'utilisateur. Son principe est de modifier la requête de l'utilisateur par ajout de termes significatifs et/ou ré-estimation de leur poids.
La dimension de l'espace de recherche étant élevée, la difficulté fondamentale de la reformulation de requête est alors la définition de l'approche à adopter en vue de réduire l'espace de recherche :
1. Critères de choix des termes de l'expansion ;
2. Règles de calcul des poids des nouveaux termes ;
3. Hypothèse de base quant aux liens entre termes et documents.
Des techniques d'amélioration des systèmes de recherches d'information (SRI) ont apparu, ces techniques ont pour objectif de guider l'utilisateur vers une bonne formulation de ses besoins. Les techniques proposées tournent autour de trois approches : [22]
- La reformulation de la requête : consiste à modifier la requête initiale de l'utilisateur par exemple, par ajout (ou suppression) des termes.
- Le ré-ordonnancement des documents : consiste soit à réordonner la liste des documents pertinents trouvés par le système sans modifier la requête initiale.
- La combinaison des résultats issus de différents SRI ou l'intégration du profil utilisateur dans le processus de recherche d'information.
Pour notre part, nous nous intéressons particulièrement à la reformulation de la requête.

Ce chapitre est organisé en deux grandes parties : la première présente les outils de base de la reformulation de requête par l'utilisation des méthodes de classification, thesaurus et les différentes techniques à utiliser pour la reformulation de requête comme la reformulation automatique basée sur le contexte global/local, oul'utilisation dela technique de reformulation par réinjection de pertinence(RF) de l'utilisateur, nous décrivons le processus général de RF et les principes d'approches à utiliser. La deuxième partie sera consacrée à la réinjection automatique de pertinence en présentant quelques travaux comme des méthodes d'extraction des termes.A la fin nous concluons le chapitre.

3.2. Reformulation de la requête :

La Reformulation de la requête est un mécanisme adaptatif de modification de requête qui a des conséquences très avantageuses sur les résultats de recherche. Cette modification de requête en poids et/ou structure peut être basée sur diverses techniques : utilisation du thesaurus, utilisation des résultats de recherche locale, injection de pertinence de l'utilisateur etc....

3.2.1. Les outils de base :

Les techniques de reformulation de requête ont généralement recours à l'utilisation de techniques de classification et du thesaurus [23].

3.2.1.1. La classification:[23]

La classification découpe l'espace des documents en sous-espaces homogènes appelés classes. Celles-ci sont constituées à partir de critères discriminatoires restreignant l'espace de recherche à un échantillon plus pertinent; les documents d'une même classe sont caractérisés par la même valeur du critère.
Plusieurs stratégies de classification ont été proposées; nous présentons brièvement les techniques basées sur les attracteurs de groupes, similarité de documents et pertinence par rapport à une requête.

a. Classification par choix d'attracteurs de groupes :

Le principe de classification consiste, dans ce cas, à choisir un document attracteur pour chaque groupe de documents. Un document est rattaché au groupe dont l'attracteur est le plus similaire, les pôles attracteurs sont déterminés préalablement par échantillonnage de documents classés par un expert. L'analyse des documents est basée sur un modèle mathématique polynomial portant sur la distribution des termes dans les documents. Le calcul de probabilité $P(D_i, C_k)$ pour que le document D_i appartient à la classe C_k est effectué selon laformule suivante :

$$P(D_j,C_k)=N_k\prod_{i=1}^{m}\frac{d_{ji}*N_k(i)+(1-d_{ji})*(N_k-N_k(i))}{N_k}$$

Où :

m : Nombre total de termes qui occurrent dans l'échantillon

C_k : kième classe

N_k : Nombre de documents de la classe C_k appartenant à l'échantillon

N_k (i) : Nombre de documents de la classe C_k appartenant à l'échantillon et contenant le terme t_i

b. Classification hiérarchique :

Cette technique est basée sur le calcul d'une matrice de similitude entre documents. Dans la stratégie de classification avec un seul passage, on construit, à partir de la matrice de similitude, un graphe de classement où les nœuds représentent des documents. Deux sommets sont reliés par une arête si le degré de ressemblance entre documents correspondants est supérieur à un seuil établi. La décomposition du graphe obtenu en classes, utilise des techniques liées à la théorie des graphes; on citera notamment les définitions suivantes :
- Une classe est une composante connexe du graphe. Une classe forme un groupe de sommets dans lequel chaque sommet est connecté à tous les autres.
- Une classe est une étoile du graphe. Une étoile est un ensemble de sommets telqu'il existe un sommet central connecté à tous les autres.
La stratégie de classement séquentiel suppose l'existence d'un critère declassification pour le critère à utiliser. On définit ainsi une hiérarchie de classesdéfinie chacune par un descripteur constitué de l'ensemble des termes d'indexationdes documents qu'elle contient. Le classement d'un document dans une classes'effectue par calcul du degré de ressemblance au

centroïde correspondant. Lecentroïde d'une classe est l'ensemble des termes représentatifs de ses documents.

c. Classification basée sur la pertinence des documents :

On associe à chaque document une coordonnée aléatoire sur un axe réel;les coordonnées des documents pertinents pour une requête sont ensuite modifiéesen vue de les rapprocher les uns des autres. Dans le but d'éviter la concentration dedocuments, le centroïde de ces documents est éloigné de celui de la collection. L'originalité de cette approche est de définir les classes de documentspertinents par fusions progressives de documents jugés pertinents mais éloignésdes requêtes en cours. Les auteurs ont mis au point des heuristiques basées sur descalculs statistiques de distribution des termes dans la collection et dans les classesafin de maintenir un équilibre entre leurs tailles.

3.2.1.2.Le thesaurus :[23]

Un thésaurus est un ensemble de mots ou syntagmes, un vocabulaire contrôlé organisé de façon à indiquer les relations entre ces mots et syntagmes et à permettre de décrire et de repérer l'information au moment voulu. Les mots et syntagmes sont ceux qui décrivent le mieux le sujet et qui, lorsqu'on établit la relation qui existe entre eux, résument le mieux l'information. Un thésaurus facilite la gestion des concepts énoncés dans un texte. Les termes sont organisés selon une hiérarchie afin d'illustrer la relation entre ce qui est général et ce qui est spécifique. On a comparé les thésaurus à des mécanismes substituts permettant de réduire l'effort intellectuel que doit normalement accomplir le chercheur. Les thésaurus peuvent être présentés de plusieurs façons (au moyen de différents graphiques utilisant, entre autres éléments, des flèches, des structures arborescentes, des modèles systématiques ou de simples listes alphabétiques). On utilisera également une variété de symboles et d'abréviations, dont les plus simples pourraient être TG pour terme générique et TS pour terme spécifique.Nous synthétisons ci dessous les principales approches adoptées pour sa construction.

a. Thesaurus manuel :

Consiste à définir interactivement divers liens linguistiques entre mots : synonymes, hyperonymes, hyponymes, polysémies etc....Les thesaurus sont organisés en catégories de mots; chaque catégorie correspond à un sens bien défini par les indexeurs. La polysémie y est traduite par la possibilité d'associer à chaque mot, *n* catégories différentes représentant ses différents sens. Ce mode de construction est généralement adapté à des collections de petites tailles, à domaine spécifique.

b. Thesaurus automatique :

Consiste à déterminer une hypothèse de liaison sémantique et l'utiliser pour la génération automatique du thesaurus. Cette liaison est généralement basée sur la cooccurrence, contexte des termes ou leur combinaison.

b.1. Thesaurus basé sur la cooccurrence :

Consiste généralement à combiner une mesure seuillée de cooccurrence entre descripteurs des termes dans la collection et un algorithme de classification, la mesure utilisée pour le calcul de la cooccurrence estla suivante :

$$SC(t_i,t_j)=\left(\frac{\sum_{k=1}^{N}\min(tf_{ik},tf_{jk})\log(\frac{N}{df_{ij}}*p_j)}{\sum_{k=1}^{N}d_{ik}}\right)*W_j$$

Où :

df_{ij} : Nombre cooccurrences entre les termes t_i et t_j

p_j : Longueur du descripteur du terme t_j

d_{ik} : Poids du terme i dans le document D_k

f_{ik} : Fréquence du terme i dans le document k

Avec :

$$W_j=\frac{\log(\frac{N}{df_j})}{\log N}$$

b.2. Thesaurus basé sur le contexte :

L'idée de base est de distinguer les polysémies par définition de contextes d'utilisation des termes dans la collection. A chaque terme est ainsi associé plusieurs vecteurs contextes dépendants de leur usage dans les documents. Dans [24], les auteurs définissent le contexte d'un terme t_l à une position voisine i, $VC_i = (W_{i1}, \ldots, W_{i,200})$ comme formé des *200* termes à plus grande valeur de cooccurrence avec le terme t à la position i,
Où :

$$W_{ik}=\log\left(\frac{N*df_{ik}}{tf_i*tf_k}+1\right)$$

Avec :

df_{lk} : Fréquence de cooccurrence de contexte du terme t_l avec le terme t_k

tf_l : Nombre total d'occurrences du terme t_l dans la collection

tf_k : Nombre total d'occurrences du terme t_k dans la collection

Le vecteur descripteur d'un terme t_l est composé de vecteurs contextes situés aux *3* positions précédentes et *3* positions successives :

$$t_l = < VC_{-1}\ VC_{-2}\ VC_{-3}\ VC_1\ VC_2\ VC_3>$$

L'utilisation d'une description contextuelle des termes est également proposée dans [31]. Dans l'approche présentée, un sens dominant est associé à chaque terme
$t_i\ (t_{i-P+1}, \ldots, t_1, \ldots, t_P)$ dans un document et représenté comme suit :
Pour chaque occurrence du terme t_i :

1. créer un contexte local constitué par P termes à droite et P termes à gauche,

2. calculer pour chaque terme t_v voisin de $t_i$dans le contexte local, le poids $Fréquence(t_v, t_i)/Fréquence(t_i)$,

3. constituer le vecteur normalisé des dix termes de plus grands poids.

Des travaux présentés ci dessous montrent l'intérêt de la représentation contextuelle des termes.

3.2.2. La reformulation automatique :

La reformulation automatique de requête induit un processus d'expansion et/ou repondération de la requête initiale en utilisant des critères de choix définis sans intervention de l'utilisateur. Ce type de reformulation peut être défini dans un contexte *global*, basé sur le thésaurus, ou alors *local*, basé sur les résultats de la recherche en cours.[24]

3.2.2.1. Reformulation basée sur le contexte global :[24]

Cette stratégie de recherche fait référence à l'exploitation d'informations préalablement établies dans la collection, et non dépendantes de la recherche en cours, en vue de réaliser la reformulation. Ceci fait alors appel essentiellement à l'utilisation de thesaurus.

a. Utilisation d'un thesaurus manuel :

Le principe fondamental est d'ajouter à la requête initiale, les termes voisins définis dans le thesaurus et sélectionnés par l'application d'un seuil et d'un algorithme de choix. L'expansion de requête basée sur l'utilisation du thesaurus manuel. La recherche d'information est effectuée selon les principales étapes suivantes :

1- Expansion de requête en utilisant les liens sémantiques prédéfinis dans le thesaurus. Plus précisément, la requête utilisateur $Q_k$est étendue avec les termes de l'ensemble défini comme suit :

$$C^* = \bigcup_{t_i \in Q_k} C_i$$

où :

$C_i = \{ t_v \in C_i / (t_v \neq 0) \wedge (C_v = C_i) \}$, C_i : Catégorie du terme t_i

En fait, on intègre à la requête utilisateur l'ensemble des termes qui traduisent la couverture sémantique de chacun de ses termes

2- Calcul de pertinence des documents selon un mécanisme d'activationpropagation basé sur le modèle connexionniste

b. Utilisation d'un thesaurus automatique basé sur la similarité :

L'expansion de requête est basée sur un thesaurus construit de façon automatique, modélisant des liens de similarité entre termes. Chaque terme t_i, s'associe un descripteur vectoriel $(d_{1i} \ldots, d_{Ni})$

Où :

$$d_{ji}=\frac{(0.5+0.5*\frac{f_{ji}}{Maxf_{ji}})}{\sqrt{\sum_{j=1}^{N}(0.5+0.5*\frac{f_{jl}}{Max_l(f_{jl})})^2 itf_j^2}}$$

Avec
f_{ji} : Fréquence du terme t_i dans le document D_j
itf_j : Fréquence inverse du document D_j

La relation entre termes est représentée par un facteur de corrélation calculé comme suit :

$$C_{UV}=\vec{t_U}\,\vec{t_V}=\sum_{j=1}^{N} d_{uj}d_{vj}$$

L'expansion de requête est alors effectuée selon les étapes suivantes :

1. Représenter sous forme vectorielle, la requête initiale

$$\vec{Q_k}=\sum_{ti\in Qk} q_{ki}\,\vec{t_i}$$

2. Utiliser le thesaurus pour calculer

$$Sim(\vec{Q_k},\vec{t_j})=\vec{Q_k}\vec{t_j}=\sum_{ti\in Qk} q_{ki}C_{ij}$$

3. Ajouter à la requête les *r* top termes $t_s$sélectionnés par *Sim* (Q_k,k_s). A chaque terme ajouté t_a, on utilise un poids donné par :

$$q_{ar}=\frac{Sim(Q_k,t_a)}{\sum_{ti\in Qk} q_{ki}}$$

c. Utilisation d'un thesaurus basé sur le contexte :

L'idée essentielle est d'étendre une requête par intégration de termes de même contexte que ceux qui la composent. Dans ce cadre, l'approche est différente au principe adopté pour la définition d'un contexte de mot. Une expansion de requête basée sur l'utilisation d'un thesaurus organisé en classes définissant des contextes. L'idée est de proposer une application d'un algorithme de classification pour l'organisation de documents; les termes d'expansion sont issus d'un thesaurus constitué des termes de faible fréquence associés à chaque requête. La sélection des termes à ajouter, est basée sur le poids de la classe calculé par la formule :

$$W_c = \frac{W_{tc}}{|C|} * 0.5$$

Où
| C | : Cardinal de la classe C
W_{tc}: Poids du terme t dans la classe C, calculé selon la formule

$$W_{tc} = \frac{\sum_{i=1}^{|C|} W_{ic}}{|C|}$$

Avec W_{ic} : Poids du terme t_i dans la classe C

L'approche proposée par Jing et E. Tzoukermann en 1999 [31] est basée sur la distance contextuelle et la proximité morphologique entre termes. La principale motivation pour l'intégration de ces deux aspects dans le modèle, est que la corrélation basée sur la morphologie d'un mot fait augmenter le rappel alors que la corrélation basée sur le sens fait augmenter la précision.

L'algorithme de recherche est effectué en deux étapes :
Etape 1 : *Préambule à la recherche*
1. Construction de la base documentaire en utilisant un analyseur morphologique
2. Construction du vecteur contexte de chaque document constitué par les vecteurs contextes de chacun de ses termes :

VC_i (t_1 (W_{i1}), ..., t_N (W_{iN}))
Où :
Wij : poids du terme ti dans le document Dj
3. Calcul des cooccurrences locales, dans la collection, pour toute paire de mots :

$$R(t_1,t_2) = \frac{IDF(t_1,t_2)}{IDF(t_1)+IDF(t_2)-IDF(t_1,t_2)}$$

Où : IDF(t_i, t_j) : Fréquence inverse de la cooccurrence des termes t_i et t_j dans la collection

Etape 2 : *Recherche*
Pour chaque requête et chaque document :
1. Calculer la distance contextuelle moyenne entre chaque terme de la requête etses variantes morphologiques dans le document selon la formule :

$$Dist(VC_1,VC_2) = \sum_{i=1}^{|VC|} R(t_i, B_m(i)) * W_{ti} W_{2m(i)}$$

Où :
|VC| : taille des vecteurs contexte

$B_{m(i)}$: terme le plus cooccurent avec t_i, ie $B_{m(i)}$ / $R(t_i, B_{m(i)}) = Max_{j=1}^{T} R(t_i,t_j)$

2. Si (distance contextuelle moyenne est supérieure à un seuil) ou (taille du vecteurcontexte est inférieur à un seuil) Alors

Considérer les deux termes équivalents *(Expansion)*
Sinon
Considérer les deux termes différents
3. Calculer la similarité requête – document
Cet algorithme permet ainsi de corréler des termes sur la base de leur morphologie mais aussi de leur sens, ceci par opposition aux algorithmes d'indexation classiques qui tronquent les termes morphologiquement reliés au même mot même s'ils nevéhiculent pas le même sens.

d. Reformulation basée sur une combinaison de thesaurus :

Plus précisément, trois types de thesaurus sont utilisés dans cette combinaison :

- Thesaurus manuel

Un terme y est représenté selon différentes taxonomies. On y associe un graphe sémantique où les nœuds représentent les termes et liens des relations de synonymie entre termes inter-taxonomies et intra-taxonomies. La similitude entre deux mots est définie comme le chemin le plus court dans le graphe :

$$Sim(t_i,t_j)=Max_{pathP}(-\log\frac{N_P}{2*D})$$

Où :

N_p : Nombre de nœuds entre t_i et t_j selon le chemin P
D : Hauteur maximale de la taxonomie

- Thesaurus basé sur la cooccurrence

On évalue la cooccurrence de pseudo-phrases de taille fixe *T* dans des blocs adjacents de documents

$$Sim(b_i,b_j)=\frac{\sum_{t=1}^{T} w_{tbi}w_{tbj}}{\sqrt{\sum_{t=1}^{T} w_{tbi}^2 \sum_{t=1}^{T} w_{tbj}^2}}$$

Où :

b_i : ième bloc

W_{tbi} : Fréquence du terme t dans le bloc bi

- Thesaurus basé sur le contexte linguistique

Les mots sont classés par contexte grammatical verbe, sujet, adjectif etc.... puis on calcule la cooccurrence relative entre deux termes, dans chaque classe, selon une formule appropriée.
Exemple : classe Adjectif

$$I(a_i,adj,n_j)=\log\frac{f_{adj}(a_i,n_j)/N_{adj}}{(f_{adj}(n_j)/N_{adj})*(f(a_i)/N_{adj})}$$

Où :

$I(a_i, adj, n_j)$: Valeur de cooccurence de a_i en qualité d'adjectif du nom n_j

$f(a_i,n_j)$: Fréquence d'occurrence de a_i en qualité d'adjectif de n_j

$f_{adj}(n_j)$: Fréquence d'occurrence de n_j en qualité d'objet de tout adjectif

N_{adj} : Nombre total d'adjectifs dans la collection

$f(a_i)$: Fréquence de l'adjectif a_i dans la collection

On calcule la similitude entre deux termes selon la formule :

$$Sim(t_i,t_j)=\frac{\sum_{(c,t)\in T(t_i)\cap T(t_j)}(I(t_i,c,t)+I(t_j,c,t))}{\sum_{(c,t)\in T(t_i)}I(t_i,c,t)+\sum_{(c,t)\in T(t_j)}I(t_j,c,t)}$$

Où :

c : Classe grammaticale (adjectif, nom, verbe …)

$T(t) = \{ (c,t') / I(t,c,t') > 0 \}$

Le principe de recherche / expansion est alors le suivant :

1. Représenter la requête sous forme vectorielle $Q(q_{k1}, \ldots, q_{kt})$

Où :

$$q_{ki}=\frac{(\log(tf_{ki})+1.0)*\log(N/n_i)}{\sqrt{\sum_{j=1}^{T}[\log(tf_{kj}+1.0)*\log(N/n_j)]^2}}$$

tf_{ki} : Fréquence d'occurrence du terme t_i dans la requête Q_k

2. Calculer la similitude entre termes de la requête et termes du thesaurus combiné comme suit :

$$Sim(Q_k,t_i)=\sum_{t_j \in Q_k} q_{kj}\overline{Sim}(t_i,t_j)$$

Où :

$\overline{Sim}(t_i,t_j)$: Similitude moyenne entre les termes t_i et t_j relativement au 3 types de thesaurus

Avec :

$\overline{Sim(t_i,t_j)}=\frac{\sum_{Typethesaurus} Sim(t_i,t_j)}{3}$, $Sim(t_i, t_j)$ est normalisée comme suit :

$$Sim(t_i,t_j)=\frac{Sim(t_i,t_j)_{old}-Sim(t_i,t_j)_{min}}{Sim(t_i,t_j)_{max}-Sim(t_i,t_j)_{min}}$$

Avec :

$Sim_{old}(t_i,t_j)$: Valeur de similititude calculée selon la formule non normalisée associée au type de thesaurus

$Sim_{min}(t_i,t_j)$: Valeur de similititude minimale calculée selon la formule non normalisée associée au type de thesaurus

$Sim_{max}(t_i,t_j)$: Valeur de similititude maximale calculée selon la formule non normalisée associée au type de thesaurus

3. Ordonner les termes par valeurs croissantes de *Sim* (Q_k, t_j). Retenir les *r* top termes pour l'expansion de requête avec un poids calculé comme suit :

$$q_{ki}=\frac{Sim(Q_k,t_i)}{\sum_{t_j \in Q_k} q_{kj}}$$

Ainsi, le poids d'un nouveau terme dépend de l'ensemble des poids de la requête mais également de la similitude relativement à chacun des types de thesaurus.

3.2.2.2. Reformulation basée sur le contexte local : [25]

Dans le cas de cette stratégie de recherche plus connue sous l'expression anglaise « ad hoc feedback », les informations utilisées pour la reformulation de requête dépendent en grande partie de la recherche en cours : documents retrouvés, termes et poids associés. A l'origine, les travaux relatifs à l'utilisation de cette stratégie consistent essentiellement en l'application de techniques de classification de termes issus des *n* tops documents retrouvés. Actuellement, de nouvelles techniques sont mises en œuvre en vue d'analyser le contexte local de la recherche et de l'exploiter pour l'expansion de requête.

L'approche proposée par [26] combine les atouts de l'analyse globale et analyse locale en procédant comme suit :

1. Identification des *n* tops passages de documents par appariement vectoriel avec larequête

2. Pour chaque concept identifié, calculer

$$Sim(Q_k,t_i)=\prod_{t_j \in Q_k}(\delta+\frac{\log(f(t_i,t_j)*idf_i}{\log(Ns)})^{idf_i}$$

Où :
Ns : Nombre de top documents sélectionnés
idf_i : Fréquence inverse du terme t_i
δ : Constante
Avec :

$$f(t_i,t_j)=\sum_{k=1}^{T} pf_{ik} * pf_{ik}$$

Pf_{jk} : Fréquence du terme t_i dans le kème passage

3. Les *r* top termes sont ajoutés à la requête avec un poids : $q_{ki}= (1-0.9*i)/m$
où i est la position du terme dans la liste

3.2.3. La reformulation par réinjection de pertinence :

La reformulation de requête par réinjection de pertinence est plus connue sous le nom de *Relevance Feedback* [28]. Cette méthode permet une modification de la requête initiale, sur la base des jugements de pertinence de l'utilisateur sur les documents restitués par le système. La relevance feedback est une forme de recherche évolutive et interactive. Son principe fondamental est d'utiliser la requête initiale pour amorcer la recherche d'information puis exploiter itérativement les jugements de pertinence de l'utilisateur afin d'ajuster la requête par expansion ou repondération. La nouvelle requête obtenue à chaque itération de feedback, permet de « corriger » la direction de recherche dans le fond documentaire, et ce, dans le sens des documents pertinents.[27]

3.2.3.1.Processus général de *RF* : [6]

Le processus de réinjection de pertinence, comme schématisé sur la figure c'est dessous, comporte principalement trois étapes : l'échantillonnage, l'extraction desévidences et la réécriture de la requête.

-L'échantillonnage : cette étape permet de construire un échantillon dedocuments à partir des éléments jugés par l'utilisateur. Cet échantillonest caractérisé par le nombre d'éléments jugés pertinents.

-L'extraction des évidences : est l'étape la plus importante, elle consiste en général à extraire les termes pertinents qui serviront à l'enrichissement de la requête initiale. Plusieurs approches ont été développées, parmi les premiers la plus reconnue est celle de *Rocchio* [28] adaptée au modèle vectoriel.

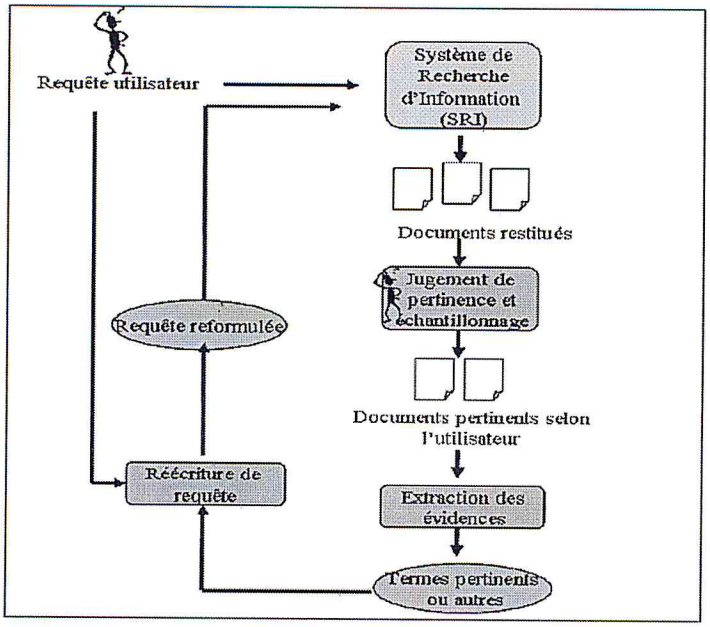

Figure 12:Le Processus général de la réinjection de pertinence [6]

-La réécriture de la requête : consiste à construire une nouvelle requête en combinant la requête initiale avec les informations extraites dans l'étape précédente.

Le processus général de la réinjection de pertinence peut être renouvèle plusieurs fois pour une même séance de recherche : on parle alors de la réinjection de pertinence à itérations multiples. [6]

Considérons maintenant en détail les différentes phases du processus de réinjectionde pertinence.
La phase d'échantillonnage ne présente pas de problématique spécifique. Le seul point aborde àce niveau concerne le nombre d'éléments à évaluer pour pouvoir effectivement constituer un échantillonreprésentatif.
La problématique principale de la réinjection de pertinence réside dans les deux autres phases: l'extraction des termes (ils sont alors pondères pour sélectionner les plus pertinents) et la réécriture de la requête avec repondération des termes.
Dans la plupart des approches de la littérature, les deux phases sont effectuées avec des méthodes de pondération des termes similaires. Cependant certainesméthodeset particulièrement celles basées sur le modèle probabiliste, utilisentdes méthodes de pondération différentes.

Dans la prochaine section nous proposons donc de détailler les méthodesreformulation de requêtes appliquée aux différents modèles de RI.

3.2.3.2. Principales approches de reformulation par réinjection de pertinence en RI :

a. Réinjection de pertinence dans le modèle vectoriel :

Dans le modèle vectoriel, la requête et les documents sont représentés sous forme vectorielle, la réinjection de pertinence consiste à rapprocher le vecteur requête à ceux des documents pertinents et l'éloigner des documents non pertinents.
La nouvelle requête Q_{i+1} est construite grâce à la **formule de Rocchio.**

a.1. Approche de Rocchio :[28]

La reformulation de requête a été introduite par Rocchio [28] dans le modèle vectoriel. La restitution des documents pertinents est liée à la notion de "requête optimale". Cette dernière est censée maximiser la différence entre le vecteur des documents pertinents et celui des documents non pertinents.
Comme l'utilisateur n'est pas en mesure de soumettre une requête optimale, la réinjection de pertinence doit permettre de rapprocher le vecteur de la requête initiale du vecteur moyen des documents pertinents et de l'éloigner du vecteur moyen des documents non pertinents. Ceci est mis en œuvre par repondération des termes initiaux et ajout de nouveaux termes pondérés à la requête initiale. Les poids servent à la discrimination des documents pertinents des documents non pertinents. La formule originale de Rocchio est définie comme suit :

$$Q_1 = Q_0 + 1/n_r \sum_{i=1}^{n_r} R_i - 1/n_s \sum_{i=1}^{n_s} S_i$$

où
Q_0 est le vecteur de la requête initiale, Q_1 est le vecteur de lanouvelle requête, $n_r$est le nombre de documents pertinents, $n_s$lenombre de documents non pertinents, $R_i$est le vecteur du *ieme* document pertinent et $S_i$le vecteur du *ieme* document non pertinent.

Le nouveau vecteur de requête est le vecteur de la requête initiale plus lestermes qui différencient au mieux les documents pertinents des documents nonpertinents. Une requêtereformulée contient de nouveaux termes (extraits desdocuments jugés pertinents) associes à de nouveaux poids. Si le poids d'unterme de la requête décroit vers zéro ou au dessous de zéro, il est éliminé del'ensemble des termes de la requête.

Une variante de cette formule a étéexaminéeexpérimentalement avec desrésultats positifs sur le système de recherche. La petite taille dela collection de documents utilisée dans les expériences de Rocchio [28] a engendrécertaines modifications dans la formule. Par exemple, un terme est seulementconsidéré s'il appartient à la requête initiale ou s'il apparait plus dans les documents pertinents que dans les documents non pertinents et dans plus quela moitié des documents pertinents. Ces modifications accentuent la difficultéd'aligner la théorie avec la pratique expérimentale.

Une autre modificationapportée à cette formule qui permet de pondérerla contribution relative de la requête initiale, des documents pertinents et desdocuments non pertinents dans le

processus de *RF*. C'est la variante la plusrépandue aujourd'hui (standard), elle est décrite dans l'équation suivante :

$$Q_1 = \alpha Q_0 + \beta / n_r \sum_{i=1}^{n_r} R_i - \gamma / n_s \sum_{i=1}^{n_s} S_i$$

où α, β, γindiquent le degré d'effet de chaque composant sur le processus de réinjection de pertinence.

b. <u>Réinjection de pertinence dans le modèle probabiliste</u> :

Dans le modèle probabiliste développé par Robertson et Sparck Jones[30], les documents et les requêtes questions sont également vu comme des vecteurs mais la mesure vectorielle de similarité est remplacée par une fonction probabiliste. On rappelle que le modèle probabiliste est basé sur la probabilité qu'un document soit pertinent à un utilisateur pour une requête donnée. Ce modèle est par essence même lié à la réinjection de pertinence, puisque ses paramètres sont estimés sur la base de la présence/absence des termes dans les documents pertinents et non pertinents.

La formule suivante utilisée pour la pondération des termes :

$$W_i = log \frac{p_i(1-q_i)}{q_i(1-p_i)}$$

W_i le poids du terme i, avec

$p_i = P(t_i = 1/D \text{ est pertinent}) = \frac{r_i}{R}$, $q_i = P(t_i = 1/D \text{ est non pertinent}) = \frac{n_i - r_i}{N - n_i}$

où $t_i = 1$ si le terme i indexe le document, $t_i = 0$ sinon.

r_i le nombre de documents pertinents contenant le terme t_i,

R le nombre de de documents pertinents pour la requête,

n_i le nombre de documents contenant le terme t_i et

N le nombre de documents dans le collection.

Les poids des termes ajoutes à la requête sont alors calcules selon la formule suivante :

$$W_i = log \frac{r_i/R - r_i}{n_i - r_i/(N - n_i) - (R - r_i)}$$

Croft [26] a défini une méthodologie de repondération en utilisant une version révisée de la formule de pondération de Robertson et Sparck Jones [30]. Plus précisément, la recherche initiale suit la fonction de pondération des termes suivante :

$$W_{ijk} = (C + idf_i).f_{ik} \qquad (2.11)$$

Avec

C une constante, f_{ik} la fréquence du terme t_i dans le document k, idf_i la fréquence absolue du terme t_i dans la collection et j la requête.

Pour ré-pondérer des termes par réinjection de pertinence, Croft se base sur la formule de Robertson et Sparck [30]. La formule de repondération est la suivante :

$$W_{ijk} = \left[C + log\frac{p_{ij}(1-q_{ij})}{q_{ij}(1-p_{ij}}\right].f_{ik}$$

W_{ijk} le poids du terme t_i dans la requête j et le document k,

$p_{ij} = \frac{r_i+0.5}{R+1.0}$ si $r_i > 0$, $p_{ij} = 0.01$ si $r_i = 0$,

$q_{ij} = \frac{n_i-r_i+0.5}{N-R+1.0}$ si $r_i > 0$, $p_{ij} = 0.01$ si $r_i = 0$,

$f_{ik} = K + (1-K).\frac{freq_{ik}}{max(freq_k)}$

où $freq_{ik}$ est la fréquence du terme t_i dans le document k, $max(freq_k)$ est le maximum des fréquences des termes dans le document k et C, K sont des constantes.

3.2.4. Réinjection automatique de pertinence :

La réinjection de pertinence décrite jusque là est basée sur les jugements de l'utilisateur. Une approche alternative, connue sous le nom de *pseudo-réinjection* ou *Blind Relevance Feedback*, est une approche pour la RF qui repose sur un échantillonnage automatique, utilise des techniques de réinjection automatique à l'aveugle pour construire une nouvelle requête [6]. Plus précisément, le systèmede recherche restitue un ensemble de documents répondant à la requête initiale.Ainsi au lieu de juger explicitement les documents, on suppose que les k premiers documents comme étant pertinents (documents pseudo-pertinents), « k » peut être entre 5 et 10 néanmoins la valeur 5 est recommandée dans [22], comme schématisé sur la figure endessous. On peut égalementconsidérer les documents qui sont restitues en fin de liste commenon pertinents. L'idée de base derrièrele pseudo réinjection de pertinence estqu'une itération de réinjectionbasée sur les documents les plus similaires à la requête initiale de l'utilisateur pourrait donner une meilleure restitution desdocuments.

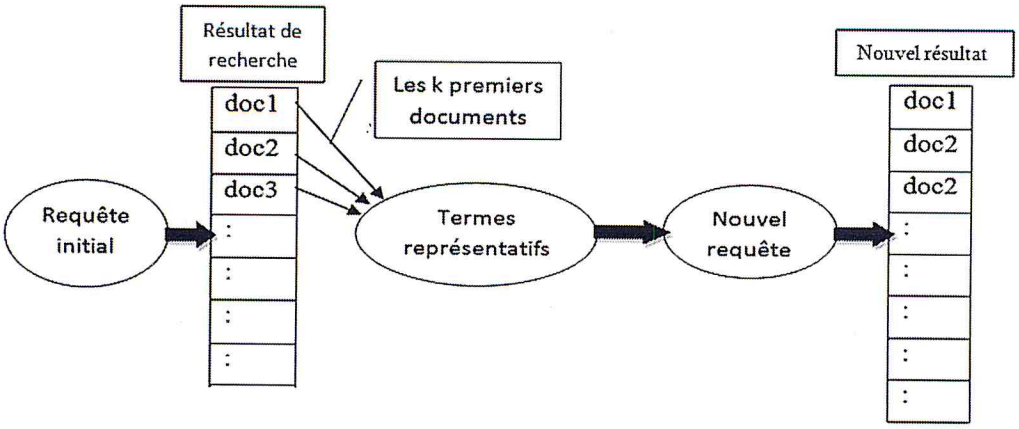

Figure 13: principe générale de réinjection automatique de pertinence

Cette technique a été développée la première fois par Rocchio[28],en tant qu'un moyen d'estimation des probabilités dans le modèle probabilistepour une première recherche. Depuis, cette technique a été largement étudiée pour améliorer les classements des documents. Croft [26] ont égalementindique que cette méthode peut avoirdes impacts négatifs. En effet si les documents considères pour la réinjection contiennent peu d'informations pertinentes ou aucune, la réinjection ajouterades termes à la requête initiale qui sont "pauvres" àdétecter la pertinence, etpar conséquent pour la recherche des documents pertinents.
La réinjection automatique peut être bénéfique si les requêtes initiales permettent de retrouver des documents pertinents, dans le cas contraire elle provoque une dégradation des performances.Des chercheurs comme Med El Amine Abderrahim et Med Alaeddine Abderrahim [22]ont essayéavecun certain succès de surmonter ce problème en améliorant le taux de précision dans les *k* meilleurs documents, c'est ce qu'on nomme habituellement la "haute précision".
Il est prouvé dans la majorité des travaux que la réinjection automatique présente une solution pratique pour l'amélioration des performances de la recherche en ligne sous un certain nombre de conditions. En particulier, c'estune technique très utile pour améliorer la recherche quand il s'agit de requêtes courtes ou de requêtes qui ne permettent pas de restituer assez de documentspertinents.

3.2.4.1. L'extraction des termes (évidence) :

L'extraction des termes ou l'extraction des évidences, est l'étape la plus importanteaprès La phase d'échantillonnage automatique, elle consiste en général à extraire les termes pertinents qui serviront à l'enrichissement de la requête initiale.

Martinet définit l'extraction des termes comme suit : l'extraction des termes est l'opération qui consiste à décrire et à caractériser un document à l'aide de représentations des concepts contenus dans ce document, c'est-à-dire à transcrire en langage documentaire les concepts après les avoir extraits du document par une analyse. [32]

Roussey définit l'extraction des termes comme étant le processus qui permet la transformation de l'information contenue dans un document vers un autre espace de représentation manipulable par un système de recherche d'information. Les objets (documents ou requêtes)

sont représentés par des descripteurs présentés sous forme d'une liste de mots clés et de poids qui leur sont associés. [33]

a. Formes d'extraction[34]

➢ Extraction contrôlée

L'extraction est un processus qui consiste à construire une représentation de chaque document au sein d'une collection ou fonds documentaire. Ce processus peut ou non être guidé par un langage documentaire (lexique de description plus ou moins organisée). S'il l'est, l'objectif est alors de choisir parmi les descripteurs préalablement définis ceux qui caractérisent au mieux le contenu du document. Ce type d'extraction est appelé : Extraction contrôlée.

➢ Extraction libre

Dans le cas contraire, lorsqu'aucun langage documentaire ne préexiste à la collection, il s'agit d'extraire de chaque document des candidats descripteurs (bien souvent des groupes nominaux), de confronter ceux-ci à ceux issus des autres documents afin d'en ériger certains au rang de descripteurs. On nomme ce type d'extraction : Extraction libre ou Extraction dérivée. Un exemple d'extraction libre est celui qui consiste à construire l'index des formes sur une collection, en éliminant les entrées peu discriminantes : Celles dont les occurrences sont fréquentes et uniformément réparties dans le document. Il en est ainsi des mots vides et de quelques autres, comme étude, méthode, etc. mais ceux-ci varient suivant les collections.

Ces deux méthodes d'indexation ont leurs avantages :

- L'extraction contrôlée assure la convergence des représentations des documents vers les seuls descripteurs autorisés. Elle réduit de ce fait la dispersion et donc le silence, contrairement à l'indexation libre.
- L'extraction libre en revanche est sensible aux évolutions de la langue, puisqu'ellen'est pas rivée par un langage préexistant. Les descripteurs peuvent évoluer au cours du temps et décrivant précisément les documents alors que l'extraction contrôlée peut laisser échapper des parties entières du contenu des documents.

b. Modes d'extraction

L'extraction permet de créer une représentation des documents dans le système. Son objectif est de trouver les concepts les plus importants du document, qui formeront le descripteur du document. L'extraction peut être [34] :
Manuelle : chaque document est analysé par un spécialiste du domaine ou par un documentaliste ;
Automatique : le processus d'extraction est entièrement informatisé ;
Semi-automatique : le choix final revient au spécialiste ou au documentaliste, qui intervient souvent pour choisir d'autres termes significatifs ;

- Extraction manuelle

L'extraction manuelle permet d'assurer une meilleure pertinence dans les réponses apportées par le SRI. Elle présente toutefois plusieurs inconvénients : deux extracteurs différents peuvent présenter des termes différents pour caractériser un même document, et un extracteur à deux moments différents peut présenter deux termes distincts pour représenter le même concept. De plus, le temps nécessaire à sa réalisation est très important.

- Extraction semi-automatique

Dans le cas d'une extraction semi-automatique, les extracteurs utilisent un thésaurus ou une base terminologique, qui est une liste organisée de descripteurs (mots clés) obéissant à des règles terminologiques propres et reliés entre eux par des relations sémantiques.

- Extraction automatique

L'extraction automatique, que nous décrivons dans ce qui suit,regroupe un ensemble de traitements automatisés sur un document. On distingue : l'extraction automatique des mots des documents, l'élimination des mots vides, la lemmatisation (radicalisation ou normalisation), le repérage de groupes de mots, la pondération des mots et enfin la création de l'index.

c. Etapes d'extraction classique

Une extraction classique, est une extraction syntaxique, qui vise la représentation d'un document textuel avec un ensemble de mot clés extrait du texte auxquels on a appliqués une normalisation. Les étapes de cette extraction sont :

➢ Extraction des termes

Un terme peut être défini comme une suite de caractères séparés par (blanc ou signe de ponctuation, caractères spéciaux,…). L'extraction des termes est le processus qui permet de convertir le texte d'un document en un ensemble de termes. Elle permet de reconnaître les espaces de séparation des mots, des chiffres, les ponctuations, etc.

➢ Elimination de mots vides

Un mot vide est un mot du langage courant et qui n'apporte pas beaucoup d'information sémantique (pronoms personnels, prépositions,...). Les mots vides peuvent aussi être des mots athématiques (les mots qui peuvent se retrouver dans n'importe quel document parce qu'ils exposent le sujet mais ne le traitent pas. On distingue deux techniques pour éliminer les mots vides :

• L'utilisation d'une liste de mots vides (stop list en anglais),
• L'élimination des mots dépassant un certain nombre d'occurrences dans la collection.

Même si l'élimination des mots vides a l'avantage évident de réduire le nombre de termesd'indexation, elle peut cependant réduire le taux de rappel, c'est à dire la proportion de documents pertinents renvoyés par le système par rapport à l'ensemble des documents pertinents. [35]

➢ Normalisation

Ce traitement consiste à retrouver pour un mot sa forme normalisée, car un mot donné peut avoir différentes formes dans un texte, mais leur sens reste le même ou très similaire. Onpeut par exemple citer économie, économiquement, économétrie, économétrique, etc. Il n'est pas forcément nécessaire d'indexer tous ces mots alors qu'un seul suffirait à représenter le concept véhiculé. Pour résoudre le problème, une substitution des termes par leur racine, ou par leur lemme (la racine dans l'exemple précédent c'est économ, par contre le lemme est économe), est utilisée. Samiha El Hamali [34]distinguent cinq types stratégiques de racinisation : la table de consultation (dictionnaire), l'élimination des affixes (on peut par exemple citer l'algorithme de Porter (Porter, 1980), la troncature, les variétés de successeurs ou encore la méthode des n-gramme.

La lemmatisation est différente à la racinisation (en Anglais Stemming). Elle consiste à retrouver pour un mot sa forme normalisée, généralement le masculin pour les noms, l'infinitif pour les verbes, le masculin singulier pour les adjectifs, etc.
Nous allons expliquer dans le paragraphe suivant quelques méthodes de normalisation :

- Elimination des suffixes

Ce sont des techniques qui permettent de donner un radical en utilisant un ensemble de règles de type : condition action, c'est-à-dire si un mot se termine par s supprimer la terminaison ou supprimer la marque du féminin, etc. Divers algorithmes de désuffixation ont été mis au point pour l'anglais, qui est une langue à morphologie relativement pauvre se prêtant donc bien à ce genre d'analyses. Les plus connus sont ceux de LOVINS, PAICE et PORTER (Porter, 1980).
L'algorithme de Porter est le plus connu est le plus utilisé, il est disponible gratuitement avec plusieurs implantation dans différents langages de programmation. C'est un algorithme basé sur des règles condition-action appliqués sur une séquence de voyelles-consonnes.

Old suffix→new suffix
Exemple:
sses→ss (caresses →caress)
ies→i (ponies→poni)
s →NULL (cats→cat)

- Troncature :

La troncature consiste à retrancher une partie du mot pour laisser un nombre déterminé de caractères représentant le radical.
Exemple : une troncature avec 7 caractères donne comme radical pour le motéconomiquement : le mot économi.
Le problème majeur avec cette méthode est le choix du nombre de caractères retenus.
Cette étude est bien approfondie par Boughanem, 2006, qu'il donne un schéma de variation du radical selon la taille de la troncature [12].

- Utilisation des n-grammes :

Un n-gram est une succession de n lettres. Généralement n = 1, 2 ou 3 lettres.
Exemple : retrieval

1. *1-gram : r, e, t, r, i, e, v, a, l
2. *2-gram : re, et, tr, ri, ie, ev, va, al
3. *3-gram : ret, etr, tri, rie, iev, eva, val

- Analyse grammaticale :

Une analyse grammaticale pour donner le radical peut se faire également avec un dictionnaire ou alors un tree-tager (logiciel gratuit sur le net).
En résumé, Quelque soit la méthode de la normalisation utilisée, cette dernière présente un certain nombre d'inconvénients qui sont: la normalisation agressive qui donne des radical qui n'ont pas de sens (universe et university donne le même radical avec l'algorithme de Porter) et l'oubli de quelques normalisation intéressantes (machine/machinery ne sont pas normalisés). La normalisation peut être également enrichie avec un traitement syntaxique des mots clés. Le premier consiste à identifier et regrouper un ensemble de mots dont la signification dépend de leur union. Par exemple, les mots « maison blanche » ne signifient

habituellement pas qu'on a affaire à une maison qui est blanche, mais plutôt au siège de la présidence des Etats-Unis. [34].

L'extraction classique appliquée dans les SRI présente plusieurs inconvénients.
Parmi ces inconvénients :
• La recherche se fait par les mots clés saisis: absence totale de la notion du « sens ».
• Cette méthode suppose que les termes soient indépendants, or ce n'est pas toujours le cas.
• Cette approche n'apporte pas de nouvelles connaissances, elle retourne les termes qui sont dans les documents tels qu'ils sont.

3.2.4.2. Pondération des termes normalisés :

***L'approche présentée par Harman en 1988** [29] consiste à sélectionner les dix premiers documents et à identifier parmi ceux-ci les documents pertinents. Harman a utilise différentes techniques pour ordonner les termes afin de choisir les vingt meilleurs termes de la liste. Il a été démontre que la technique utilisée pour le tri des termes pertinents a un large impact sur la performance. Dans plusieurs techniques de tri que l'auteur a définies, il utilise une mesure de bruit n_k calculée comme suit :

$$nk = \sum_{i=1}^{N} \frac{tf_{ik}}{f_k} log2 \frac{f_k}{tf_{ik}} \qquad 2.1$$

Avec :
tf_{ik} le nombre d'apparition du terme k dans le document i,
f_k le nombre d'apparition du terme k dans la collection et
N le nombre de termes dans la collection.

La technique a été étendue pour tenir compte du nombre de documents dans l'ensemble des documents pertinents contenant le terme k (p_k) et du nombre d'apparition du terme k dans l'ensemble des documents pertinents (rtf_k).
Harman a défini ainsi une autre mesure de bruit par rapport à l'ensemble des documents pertinents. Cette mesure est calculée comme suit :

$$rn_k = \sum_{i=1}^{N} p_k \frac{tf_{ik}}{rtf_k} log_2 \frac{f_k}{tf_{ik}} 2.2$$

Harman a défini d'autres techniques de tri des termes. La technique qui conduit à de meilleurs résultats est basée sur une formule de pondération :

$$W_{ij} = log_2 \frac{p_{ij}(1-q_{ij})}{q_{ij}(1-p_{ij})} 2.3$$

Avec :

Wij poids du terme i dans la requête j,

p_{ij} la probabilité que le terme i apparaisse dans les documents pertinents pour la requête j,

q_{ij} la probabilité que le terme i apparaisse dans les documents non pertinents pour la requête j,

La sélection des termes ayant une valeur de poids importante revient à sélectionner les termes caractéristiques des documents pertinents avec une faible probabilité d'apparition dans les documents non pertinents.
Harman a également démontré que la meilleure méthode de sélection des termes issus des documents pertinents devient inefficace au-delà de 20 à 40 termes ajoutes.

Croft [26] est adoptée une méthode de sélection de nouveaux termes sur la base d'une fonction qui consiste à attribuer à chaque terme un nombre traduisant sa valeur. La formule suivante pour calculer la valeur de sélection d'un terme :

$$selValue(i) = W_{ij} * (P_i - U_i) \qquad 2.4$$

Avec :

W_{ij} défini dans l'équation 2.3,

P_i la probabilité (d_i = 1/D est pertinent),

U_i la probabilité (d_i = 0/D est non pertinent).

Les termes sont alors tries en fonction de leurs valeurs de pertinence puis sélectionnés en utilisant un seuil prédéfini

***L'approche présentée par Med El Amine Abderrahim et Med Alaeddine Abderrahim en 2012** **[22]** consiste à sélectionner les cinq premiers documents résultant de la requête initiale sont supposés pertinents et par conséquent, on extrait les termes les plus pertinents issus de ces documents dans la phase d'extraction des évidences. Il faut noter toutefois que cette approche ne peut être bénéfique que si la requête initiale permet de retrouver des documents pertinents, sinon elle peut avoir pour effet l'augmentation de la « dérive » de la requête.
Les essais effectués dans le cadre de la compagne d'évaluation montrent que la technique de RFA permet des gains en précision des réponses d'environ 10%.Pour l'expérimentation de l'approche, ils ont utilisé un corpus de textes Arabe de différents domaines.En plus, ils ont développé un jeu de 50 requêtes dans les différents domaines.

Med El Amine Abderrahim et Med Alaeddine Abderrahimont défini d'autres techniques pour extrait les termes. La technique qui conduit à de meilleurs résultats est basée sur une formule (voir la figure c'est dessous) :
-Construire la matrice d'association locale (terme-terme) qui quantifie les relations de corrélation entre les termes issus de l'ensemble des documents retournés en réponse à la requête initiale. Selon la méthode de construction des relations de corrélationde l'ensemble des termes distincts de D_p: S avec chaque élément

$$S_{u,v} = \sum_{d_j \in D_p} f_{S_{u,j}} \times f_{S_{v,j}}$$

$f_{S_{u,j}}$: représente la fréquence du terme S_u dans le document d_j

$S_{u,v}$: exprime la corrélation entre le terme u et v

-Ils ont également démontré que la meilleure méthode de sélection des termes issus des documents pertinents. Pour chaque terme t de $q_i$extraire son cluster d'association locale C_i formé des plus grandes valeurs $S_{u,}$ $(v \neq u)$ de la $u^{ème}$ligne de S

Début // **Algorithme RI avec RFA**
Pour chaque requête q_i $(i=1,50)$

1- Interrogation de la collection des documents
2- ***Echantillonnage*** : Sélectionner les 'p' premiers documents retournés $(p=5)$: D_p (Nous avons fixé p à 5 documents, généralement p est entre 5 et 10)
3- ***Extraction des évidences***
   - Construire la matrice d'association locale (terme-terme) de l'ensemble des termes distincts de D_p : $\vec{S}$ avec chaque élément
     $$S_{u,v} = \sum_{d_j \in D_p} f_{S_{u,j}} \times f_{S_{v,j}}$$
     $f_{S_{u,j}}$: représente la fréquence du terme S_u dans le document d_j
     $S_{u,v}$: exprime la corrélation entre le terme u et v
   - Pour chaque terme t de q_i extraire son cluster d'association locale C_i formé des plus grandes valeurs $S_{u,v}$ $(v \neq u)$ de la $u^{ème}$ ligne de $\vec{S}$
4- ***Réécriture de la requête*** : Construire la nouvelle requête $q_{new} = q_i \cup C_i$

Fin

Figure 14: Algorithme de RFA par [22].

Les résultats obtenus montrent qu'il y a une augmentation en nombre de documents retournés après la RFA dans 35 requêtes (soit 70% des requêtes) et une augmentation en nombre de documents pertinents retournés après la RFA(Relevance Feedback automatic) dans 29 requêtes (soit 58% des requêtes).

3.3. Conclusion :

Nous avons présenté dans ce chapitre les outils de base et les différentes techniques à utiliser pour la reformulation de requête comme la reformulation automatique, reformulation par réinjection de pertinence(RF) de l'utilisateur et réinjection automatique de pertinence,nous avons présentéquelques travaux sur les méthodes d'extraction des termes.

Dans notre étude, nous proposons d'évaluer la technique de réinjection de la pertinence automatique pour les traitements textuels. Nous avons commencé par décrire la technique de réinjection de la pertinenceen intégrant le contexte utilisateur, ensuite, nous avons exposé notre expérimentation avec une discussion des résultats obtenus.

Chapitre 4

Réinjection automatique de la pertinence dans une recherche d'information contextuelle

4.1 Introduction :

Les SRI sont des outils informatiques qui ont pour but la mise en relation des informations contenues dans le corpus documentaire d'une part, et les besoins de l'utilisateur d'autre part. Le défit est de pouvoir retourner le maximum de documents pertinents tout en limitant le bruit et le silence documentaire, c'est-à-dire de ne pas restituer des documents qui ne répondent pas au besoin, et en même temps restituer le maximum de documents pertinents.

Dans ce chapitre nous présentons une contribution traduisant un point de vue relatif à l'amélioration des systèmes de recherche d'information. La contribution présentée s'inscrit dans le domaine de la réinjection automatique de la pertinence dans la recherche contextuelle d'information sur le web et propose une nouvelle approche basée sur les profils utilisateurs et la réinjection automatique de la pertinence pour la reformulation des requêtes. Cette approche est divisée en deux phases principales. La première phase sert à trouver dix (10) documents pertinents selon le profil utilisateur, ces documents sont utilisés dans la deuxième phase qui est la réinjection automatique de la pertinence.

Ce chapitre est organisé en deux grandes parties reflétant les deux phases. On va commencer par une architecture générale de notre contribution, ensuite on va parler dans la partie une sur la première phase, la construction du profil utilisateur et les différentes méthodes de sélection des documents pertinents selon ce profil. Après ça on va parler dans la partie deux sur la deuxième phase d'intégration la réinjection automatique de pertinence comme une modèle de reformulation de requête et aussi la méthode de choix des termes à ajouter a la requête initial a partir le résultat de la première phase.

4.2. La dimension du contexte choisi dans notre travail :

Comme on a vu dans la section 2.2.2 de chapitre 2 plusieurs taxonomies de contexte ont apparu dans la littérature, ces taxonomies sont basées sur le concept du contexte multidimensionnel. Parmi ces taxonomies on trouve la taxonomie de Tamine et al qui est la plus connue, la plus récente et la plus utilisée. Cette taxonomie comprend cinq dimensions principales : dispositif d'accès à l'information, contexte spatio-temporel, contexte de l'utilisateur, contexte de la tâche et contexte du document.

Dans notre travail on a choisis le contexte de l'utilisateur (profil utilisateur), car c'est la dimension principale abordée par la communauté [11], et la plus explorée en RI contextuelle [1]. Cette dimension permet de modéliser et de stocker les données caractérisant l'utilisateur.

4.3. Architecture générale du système:

L'architecture générale de notre système est basée sur la modélisation et l'exploitation d'un profil de l'utilisateur décrivant ses centres d'intérêts pour trouver des documents pertinents (les k premiers documents) qui sont utilisés dans la réinjection automatique de la pertinence.

Le but de l'utilisation du profil utilisateur est d'éviter le problème de l'augmentation de la « dérive » de la requête dans la réinjection automatique de la pertinence, c'est-à-dire cette dernière ne peut être bénéfique que si la requête initiale permet de retourner des documents pertinents [22], sinon on va tomber dans ce problème de la dérive de la requête, et pour cette raison on a intégrer le profil utilisateur (contexte utilisateur) pour assurer d'avoir des documents pertinents et pour garder aussi l'idée générale de la réinjection automatique de la pertinence, c'est-à-dire avoir des documents pertinents mais sans l'intervention directe de l'utilisateur (automatique).

Donc l'architecture générale de notre contribution est composée de six (6) étapes principales qui sont comme suit :

- **Etape 1 :** l'utilisateur exprime son besoin par une requête (requête initiale) sur notre application.
- **Etape 2 :** notre application va reformuler la requête initiale selon le profil utilisateur ensuite elle envoie la requête reformuler au moteur de recherche (Google).
- **Etape 3 :** le moteur de recherche retourne les k premier résultats à notre application.
- **Etape 4 :** l'extraction des termes à partir des k documents renvoyés par le moteur de recherche.
- **Etape 5 :** l'utilisation de l'algorithme de figure 16 proposer par [22] avec amélioration a pour but avoir les termes à ajouter à la requête initiale d'une part et enrichissement du profil utilisateur d'un autre part.
- **Etape 6 :** afficher le résultat final à l'utilisateur à partir du moteur de recherche Google

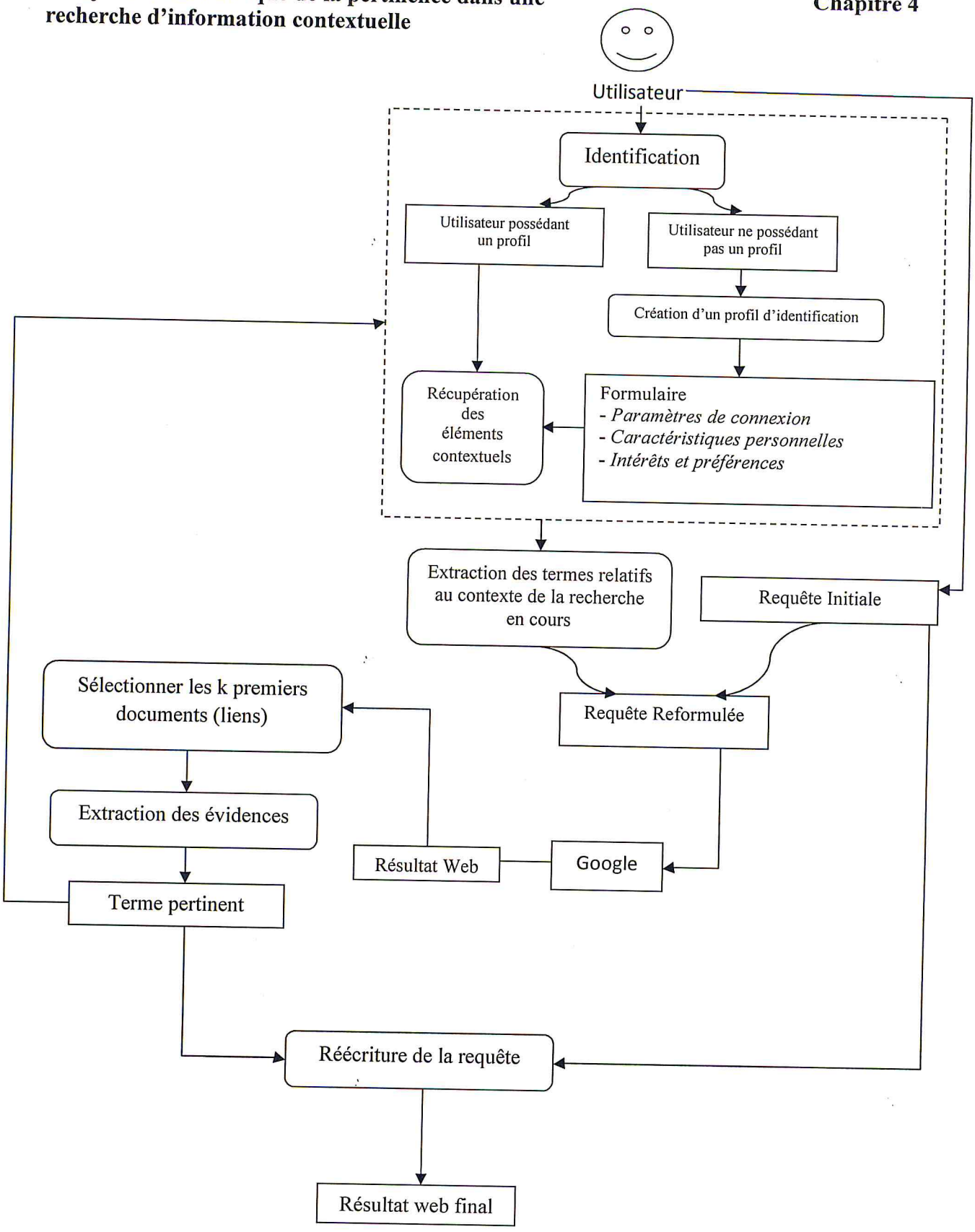

Figure 15 : Architecture générale du système

4.4. Utilisation du profil utilisateur pour trouver les documents pertinents :

La modélisation et l'exploitation du profil dans le processus de recherche consiste à trouver les documents pertinents pour l'utilisateur sans l'intervention directe de ce dernier, et cela se fait par la connaissance de ses domaines d'intérêts et ses préférences. Plusieurs approches de modélisation et d'exploitation du profil sont apparus, et on va les représenter dans ce que suit on notant l'approche qu'on va l'utiliser.

4.4.1. Modélisation du profil utilisateur :

La modélisation du profil utilisateur consiste à trouver la bonne manière de représenter l'utilisateur. Plusieurs approches ont apparu, et pour la majorité de ces approches, le profil contient la description du centre d'intérêt de l'utilisateur [chapitre2- partie2.5.1.3].

4.4.1.1. L'approche de modélisation du profil utilisateur choisi dans notre travail :

Les utilisateurs sont divers et complexes : ils sont caractérisés par des modèles cognitifs différents et font partie d'une communauté. Ils effectuent des tâches multiples ayant des buts différents. Ils ont également des activités simultanées de recherche, interactives et connexes à d'autres entités dans un domaine donné. On constate ainsi que les informations caractérisant un utilisateur ne sont pas factuelles mais multidisciplinaire. Cependant, cette diversité n'est généralement pas fidèlement représentée par les vecteurs de mots clés et les ontologies contrairement à la modélisation multidimensionnelle [4]. Et pour cette raison on a choisis dans notre travaille la modélisation multidimensionnelle, parce qu'elle permet de sécuriser le profil utilisateur [4], en plus de ça chaque profil est associé à un seul utilisateur grâce aux autres dimensions, par contre dans les ontologies et les vecteurs de mots clés qui contiennent une seule dimension (centre d'intérêts).

4.4.2. L'acquisition des données utilisateurs :

L'acquisition des données utilisateurs permet de collecter les informations des utilisateurs et ces centres d'intérêts, plusieurs approches ont apparu pour l' acquisition de profil utilisateur [chapitre2- partie2.5.1.4].

4.4.2.1. L'approche d'acquisition choisi dans notre travail :

Pour notre travaille on va utiliser les deux approches. Donc on va créer un profil statique qui contient les informations personnelles, identité, centre d'intérêts … etc de l'utilisateur puisque on a choisi la modélisation multidimensionnelle, ces informations sont introduites par l'utilisateur, et puisque l'acquisition explicite est souvent utilisé pour déclencher la collecte implicite d'informations [4], donc on va créer un autre profil dynamique, le système récupère automatiquement des informations qu'on suppose pertinentes pour enrichir le profil dynamique et les proposent à l'utilisateur. Ce dernier valide par la suite celles qu'il juge réellement pertinentes parmi l'ensemble de propositions.

4.4.3. Exploitation du profil utilisateur :

La personnalisation du processus d'accès à l'information consiste à intégrer ou exploiter le profil utilisateur dans la chaîne d'accès à l'information. Le but fondamental des modèles d'accès personnalisé à l'information est de restituer en haut de la liste des résultats des documents qui intéressent l'utilisateur dans sa recherche, en d'autres termes qui semblent les plus similaires à son profil [3], plusieurs approches ont apparu pour l'exploitation de profil utilisateur [chapitre2- partie2.5.1.5].

4.4.3.1. L'approche d'exploitation du profil utilisateur choisi dans notre travail :

D'après ce qu'on a vu précédemment et les recherches qui ont faites, on trouve que la reformulation des requêtes est la plus utilisée parce que le fait d'enrichir la requête de l'utilisateur avec les informations de son profil permet de cibler mieux les informations dont il a réellement besoin [2]. Par contre on trouve pour l'approche d'appariement documents-profils, qu'il existe peu de systèmes qui fournissent réellement une telle approche de personnalisation [4]. Et pour l'approche de ré-ordonnancement des résultats par le profil, on peut tomber dans le problème de si les résultats ne sont pas vraiment pertinentes parce que le profil ici est utilisé à la fin de la recherche (post recherche). Pour ces raisons là on a choisis l'approche basée sur la reformulation des requêtes pour trouver les k premiers documents pertinents qui seront utilisés dans la réinjection automatique de la pertinence.

4.4.4. Architecture de notre travail pour l'utilisation du profil utilisateur :

L'Architecture de notre travaille pour l'utilisation du profil utilisateur est composée de trois modules pour trouver les k premiers documents pertinents qui seront utilisés dans la deuxième phase qui est la réinjection automatique de pertinence. Il s'agit dans un premier temps de capturer les deux types de contexte nécessaires à la catégorisation de l'utilisateur (profil statique et dynamique), puis de les utilisés par le module de reformulation pour générer une nouvelle requête à partir de la requête initiale.

4.4.4.1. Module pour la capture du contexte statique :

Cette première composante du contexte sert à identifier un utilisateur à travers une série d'informations afin de le catégoriser. Le contexte statique est défini à la première connexion au système, à cet effet nous avons défini quatre catégories d'informations relatives au contexte statique, ces informations se résument en :

- Les paramètres de connexion : e-mail, mot de passe.
- Les caractéristiques personnelles : nom, prénom, pays, langue primaire, langue secondaire.
- Les intérêts et préférences : domaine, domaine secondaire, spécialité.

4.4.4.2. Module pour la capture du contexte dynamique :

Dans le but d'optimiser la réutilisation des profils et faciliter leur compréhension, cette deuxième composante du contexte consiste en l'association de la recherche au contexte de l'utilisateur. A la fin de chaque session de recherche le module de capture du contexte dynamique procède à l'extraction automatique d'un ensemble d'éléments relatifs au contexte de l'utilisateur, il les propose à l'utilisateur. Ce dernier valide par la suite ceux qu'il juge réellement pertinents. Ces informations seront enfin stockées dans la base des contextes utilisateurs.

4.4.4.3. Module de reformulation :

Le module de reformulation a pour objectif de produire une nouvelle requête à partir de la requête initialement formulée par l'utilisateur et cela en rajoutant des termes issus de son contexte de recherche actuelle. Dans un premier temps l'utilisateur formule sa requête en utilisant ses propres termes, par la suite le système procède à l'extraction de l'ensemble des termes à rajouter afin de produire une nouvelle requête, ces termes sont extraits de la base des contextes utilisateurs. Une fois la requête reformulée elle sera envoyée au moteur de recherche.

4.5. La réinjection automatique de pertinence

Dans notre approche, on propose la réinjection automatique de pertinence comme une approche de reformulation de requête qui consiste à modifier la requête initiale sans l'intervention de l'utilisateur. A partir la section 4.2.4 de chapitre 04, le processus de réinjection de pertinence, comporte principalement trois étapes : l'échantillonnage automatique, l'extraction des évidences et la réécriture de la requête.

4.5.1. L'échantillonnage automatique :

Cette étape permet de construire un échantillon de documents à partir des éléments jugés par le profil. Cet échantillon est caractérisé par le nombre d'éléments jugés pertinents.

Pour respecter l'idée général de la réinjection automatique et éviter la lourdeur de l'opération de jugement de pertinence, on extraire les 'k' premiers documents pertinent à partir les résultats 'n' de requête initialement restitués de l'étape 01, on se base sur les résultats du recherche faite par [7].

Pour notre expérimentation nous allons fixer 'k' par 5, 7, 10, c'est ainsi que, les 'k' premiers documents résultant de la requête initiale sont pertinents après la reformulation par le profil utilisateur et par conséquent, on extrait les termes les plus pertinents issus de ces documents (k premiers) dans la phase d'extraction des évidences. Et noter que cette approche est bénéfique après intégration le contexte utilisateur que la requête initiale permet de retrouver des documents pertinents.

4.5.2. Extraction des évidences :

Dans notre l'aire travail sur les corpus web et plus précisément dans l'étape prétraitement du donnée, sont divisé en deux, la première qui demande une extraction contrôlée pour enlever les informations non pertinentes du corpus comme les balises HTML (h3, p, a, li, …), PDF et doc(x) (<</D[109 0 R/Fit]/S/>>, hچق—énغ8_€ب ïبإ]_, ...) donc le nettoyage consiste à identifier et nettoyer le bruit. Et pour la deuxième, l'extraction utilisée est l'extraction libre, car les descripteurs seront des termes des documents (qui seront par la suite filtrés), et non prédéfinis à l'avance. On a utilisé l'extraction automatique comme utilisation liste des mots vides (stop list en anglais) a l'avantage évident de réduire le nombre des termes indexer, elle peut cependant réduire la quantité des informations trouvées, c'est à dire la proportion de documents pertinents renvoyés par le système par rapport à l'ensemble des documents pertinents.

Et en suite dans l'étape de pondération des termes normalisés consistent à associer à chaque terme un poids en appliquant la méthode vu dans la partie 3.2.4.2 du chapitre précédent avec une amélioration, qui consiste à mesurer la fréquence d'apparition des mots de la requête initiale et de documents (k premiers) dans la liste des mots normaliser. Dans notre approche on construit la matrice d'association locale (terme(u)-terme(v)) de l'ensemble des termes distincts de requête et *'k'* documents: *S* avec chaque élément

$$S_{u,v} = \sum_{dj \in k} \int S_{u,j} \times \int S_{v,j}$$

$\int S_{u,j}$: représente la fréquence du terme de la requête initial de l'utilisateur S_u dans le document d_j avec j= 1…k. (Le bute est minimisée la matrice d'association para port de proposition de ***[Med El Amine Abderrahim and Med Alaeddine Abderrahim]***)

$\int S_{v,j}$: représente la fréquence du terme du 'k' documents S_v dans le document d_j avec j= 1…k.

$S_{u,v}$: exprime la corrélation entre le terme *u* de requête et *v* de documents

Voir le figure 22 qu'on met dans la matrice chaque documents qui appartient à 'k' documents pertinent la valeur de $S_{U,V}$.

Trm_V / Trm_U	Trm_1	Trm_2	Trm_3	Trm_4	……	Trm_n
$Trme_1$	$S_{1,1}$	$S_{1,2}$	$S_{1,3}$	Nul	……	$S_{1,n}$
$Trme_2$	$S_{2,1}$	Nul	$S_{2,3}$	$S_{2,4}$	……	$S_{2,n}$
$Trme_3$	$S_{3,1}$	$S_{3,2}$	$S_{3,3}$	$S_{3,4}$	……	Nul
Trm_4	Nul	$S_{4,2}$	$S_{4,3}$	$S_{4,4}$	……	$S_{4,n}$
……	……	……	……	……	……	……
$Trme_a$	$S_{a,1}$	$S_{a,2}$	$S_{a,3}$	$S_{a,4}$	……	$S_{a,n}$

Figure 16: corrélation entre le terme *u* et *v* à partir du 'k' documents

Avec :

Trme$_U$ représente les termes de la requête initiale,
Trm$_V$ représente les termes des k documents,
'a' nombre totale des termes dans la requêtes initial q_i,
'n' nombre totale des termes dans les 'k' premiers documents pertinents,
Mettre nul si terme de la requête initial égale le terme des k documents pertinents.

Et ensuite dans l'étape de traitement des données, une fois le texte transformé en une représentation mathématique, on peut le classer parmi un ensemble d'autres textes a pour bute de choisir les termes pertinent à ajouter a la requête initiale donc :

Pour chaque terme '*t*' de requête initial q_i extraire son cluster d'association locale C_i formé des plus grandes valeurs $S_{u,v}$ ($v \neq u$) de la *u-Emme* ligne de *S*
C'est-à-dire :

* On cherche dans chaque colonne le S_v le plus fréquent corrélée entre le terme '*u*' de la requête et le terme '*v*' de 'K' documents donc *la Maximisation*
* et ensuite on ajoute cette terme au cluster d'association local C_i sans réputer le terme dans celle-ci.

4.5.3. Réécriture de la requête

Qui consiste à construire une nouvelle requête en combinant la requête initiale avec les informations extraites dans l'étape précédente. Le nouveau vecteur de requête est le vecteur de la requête initiale plus les termes qui différencient au mieux les documents pertinents des documents non pertinents. Une requête reformulée contient de nouveaux termes (extraits des documents jugés pertinents)

$$q_{reformuler} = q_i \cup C_i$$

Il reste juste d'envoyer la requête reformulé au moteur de recherche pour présenter le résultat final.

4.6. Conclusion :

Pour répondre à l'un des objectifs de notre travail qui consiste à exploiter utilisation de la réinjection automatique de pertinence à pour bute de renvoyer des documents plus pertinents à l'utilisateur, nous avons présenté une approche qui consiste à intégrer le profil utilisateur dans le processus de la réinjection automatique de pertinence pour assurer que l'échantillon automatique de documents est pertinents et éviter le problème de la dérivé de la requête.

On a parlé au début sur la manière qu'on a choisis pour modéliser le profil et l'intégration de ce dernier, l'idée était de reformuler la requête initiale par enrichissement, les termes ajouter à la requête sont issu à partir du profil utilisateur, cette idée nous a permet d'avoir un ensemble de documents pertinents selon le profil utilisateur et sans l'intervention de l'utilisateur. Et pour la deuxième on a choisie la réinjection automatique de pertinence comme une méthode de reformulation de requête initiale a partir de l'échantillon de document retourné de l'étape 01 à pour bute de sélectionner un ensemble des termes pour ajouter a la requête initiale. Le résultat final de cette approche permet de envoyer des documents plus pertinents à l'utilisateur.

Chapitre 05

Implémentation et évaluation

5.1. Introduction :

Dans ce chapitre, nous présentons l'environnement de développement de l'application et son évaluation. Dans ce qui suit, nous montrons le langage de programmation choisi et les outils utilisés. Nous expliquons ensuite le fonctionnement général de notre application et nous terminons par son évaluation.

5.2. Outils et environnement de développement :

Le langage choisi pour le développement de l'application est le langage JAVA, avec utilisation de l'IDE Eclipse. Nous avons exploité le kit de développement Wamp notamment son SGBDR intégré MySql pour la gestion de la base de données.

5.2.1. Langage JAVA :[25]

JAVA est à la fois un langage de programmation informatique orienté objet et un environnement d'exécution informatique portable créé par James Gosling et Patrick Naughton employés de Sun Microsystems avec le soutien de Bill Joy (cofondateur de Sun Microsystems en 1982), présenté officiellement le 23 mai 1995 au SunWorld.

Java est à la fois un langage de programmation et un environnement d'exécution. Le langage Java a la particularité principale que les logiciels écrits avec ce dernier sont très facilement portables sur plusieurs systèmes d'exploitation tels que Unix, Microsoft Windows, Mac OS ou Linux avec peu ou pas de modifications... C'est la plate-forme qui garantit la portabilité des applications développées en Java.

Le langage reprend en grande partie la syntaxe du langage C++, très utilisé par les informaticiens. Néanmoins, Java a été épuré des concepts les plus subtils du C++ et à la fois les plus déroutants, tels que l'héritage multiple remplacé par l'implémentation des interfaces. Les concepteurs ont privilégié l'approche orientée objet de sorte qu'en Java, tout est objet à l'exception des types primitifs (nombres entiers, nombres à virgule flottante, etc.).

Java permet de développer des applications autonomes mais aussi, et surtout, des applications client-serveur. Côté client, les applets sont à l'origine de la notoriété du langage. C'est surtout côté serveur que Java s'est imposé dans le milieu de l'entreprise grâce aux servlets, le pendant serveur des applets, et plus récemment les JSP (JavaServer Pages) qui peuvent se substituer à PHP, ASP et ASP.NET.

Les applications Java peuvent être exécutées sur tous les systèmes d'exploitation pour lesquels a été développée une plate-forme Java, dont le nom technique est JRE (Java Runtime Environment - Environnement d'exécution Java). Cette dernière est constituée d'une JVM (Java Virtual Machine - Machine Virtuelle Java), le programme qui interprète le code Java et le convertit en code natif. Mais le JRE est surtout constitué d'une bibliothèque standard à partir de laquelle doivent être développés tous les programmes en Java. C'est la garantie de

portabilité qui a fait la réussite de Java dans les architectures client-serveur en facilitant la migration entre serveurs, très difficile pour les gros systèmes.

5.2.2. Eclipse : [36]

Eclipse IDE est un environnement de développement intégré libre (le terme *Eclipse* désigne également le projet correspondant, lancé par IBM) extensible, universel et polyvalent, permettant potentiellement de créer des projets de développement mettant en œuvre n'importe quel langage de programmation. Eclipse IDE est principalement écrit en Java (à l'aide de la bibliothèque graphique SWT, d'IBM), et ce langage, grâce à des bibliothèques spécifiques, est également utilisé pour écrire des extensions.

La spécificité d'Eclipse IDE vient du fait de son architecture totalement développée autour de la notion de plug-in (en conformité avec la norme OSGi) : toutes les fonctionnalités de cet atelier logiciel sont développées en tant que plug-in.

Plusieurs logiciels commerciaux sont basés sur ce logiciel libre, comme par exemple IBM Lotus Notes 8, IBM Symphony ou Websphere Studio Application Developer.

5.2.3. MySql :[37]

Système de gestion de bases de données relationnelles (SGBDR) sous licence GNU (open source gratuit) très utilisé pour mettre en ligne des bases de données. MySQL fonctionne sur beaucoup de plates-formes différentes, incluant AIX, BSDi, FreeBSD, HP-UX, Linux, Mac OS X, NetBSD, OpenBSD, OS/2 Warp, SGI Irix, Solaris, SunOS, SCO OpenServer, SCO UnixWare, Tru64 Unix, Windows 95, 98, NT, 2000 et XP.

5.2.4. Wamp :[38]

Wampserver est ce qu'on pourrait appeler un serveur web local pour windows. Par abus de langage, Wampserver (auparavant nommé WAMP5) est souvent désigné par WAMP.

Acronyme signifiant « **W**indows **A**pache **M**ySQL **P**HP (dans la majorité des cas mais aussi parfois, « **P**erl » ou « **P**ython ») », il comprend un programme destiné à se comporter comme un serveur web sur votre ordinateur.

Lorsque vous demandez à votre navigateur (comme Firefox, IE8, Chrome ou Opera par exemple) d'afficher une page web, celui-ci envoie une requête au serveur possédant cette page qui le lui envoie. Wampserver se comporte exactement de la même manière sauf qu'il se trouve directement sur votre machine. Il y a donc aucune information transmise sur l'extérieur et vous pouvez donc tester votre site sans même avoir un hébergement ni même accès à internet. C'est comme s'il se trouvait en ligne sur le web.

Il comprend la suite Apache, MySQL et PHP.

Les rôles de ces quatre composants sont les suivants :

- Apache est le serveur web « frontal » : il est « devant » tous les autres et répond directement aux requêtes du client web

(navigateur)
- Le langage de script PHP sert la logique et permet des traitements (calculs, vérification, test, etc.)
- MySQL stocke toutes les données de l'application (c'est une base de données)
- Windows assure l'attribution des ressources à ces trois composants.

Il existe de la même manière son homologue sous linux désigné par LAMP ou XAMPP (le X correspondant à Unix) et sous Mac Os avec MAMP.

Ce logiciel est donc parfait pour tous les développeurs de sites web qui aimeraient tester leurs créations avant leur mise en ligne, tester des scripts mais aussi définir un site disponible sur le réseau local comme on en trouve dans beaucoup de grandes entreprises pour tenir au courant le personnel.

5.3. Explication du fonctionnement de l'application :

5.3.1. Intégration du profil utilisateur :

L'application communique avec une base de données pour stocker les éléments contextuels. Cette base de données contient cinq tables, la première sert à garder les paramètres de connexion, la deuxième pour garder les caractéristiques personnelles, la troisième pour garder les domaines de préférences, la quatrième pour garder les domaines secondaires et la cinquième pour garder les spécialités.

5.3.1.1 Fonctionnement générale :

Le fonctionnement général de notre application est comme suit : en premier temps l'utilisateur se connecte à l'application. Si ce dernier n'a pas déjà remplis ses informations (récupération du contexte statique) (Figure 17, Zone A), il doit s'inscrire pour récupérer ses derniers (Figure 17, Zone B) à travers un wizard (Figures 18). Après la connexion l'utilisateur doit choisir le domaine de recherche (Figure 19, Zone A) ensuite il exprime sa requête (Figure 19, Zone B). Après ça, l'application reformule la requête initiale de l'utilisateur suivant son profil et elle l'envoie au moteur de recherche Google.

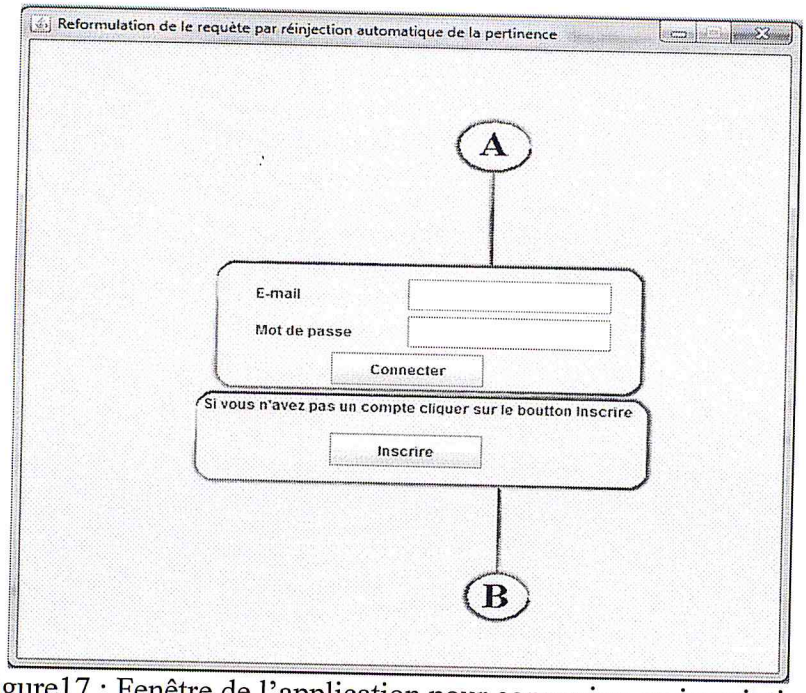

Figure17 : Fenêtre de l'application pour connexion ou inscription

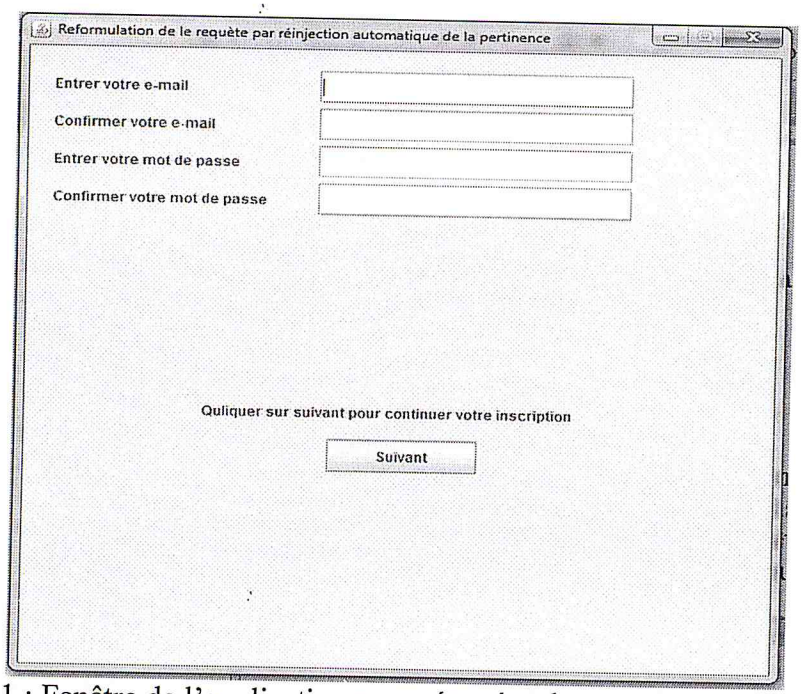

Figure 18.1 : Fenêtre de l'application pour récupérer les paramètres de connexion

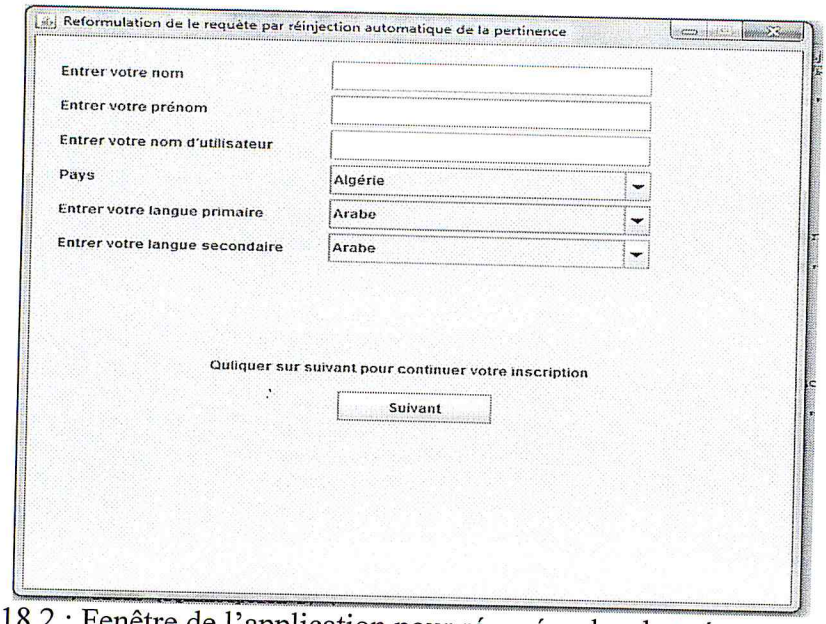

Figure 18.2 : Fenêtre de l'application pour récupérer les données personnelles

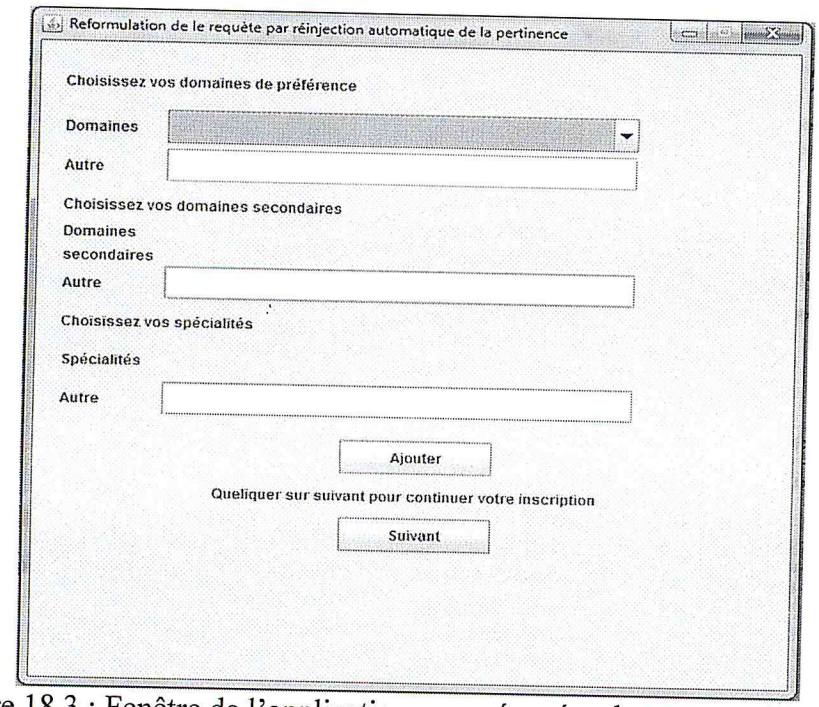

Figure 18.3 : Fenêtre de l'application pour récupérer les centres d'intérêts

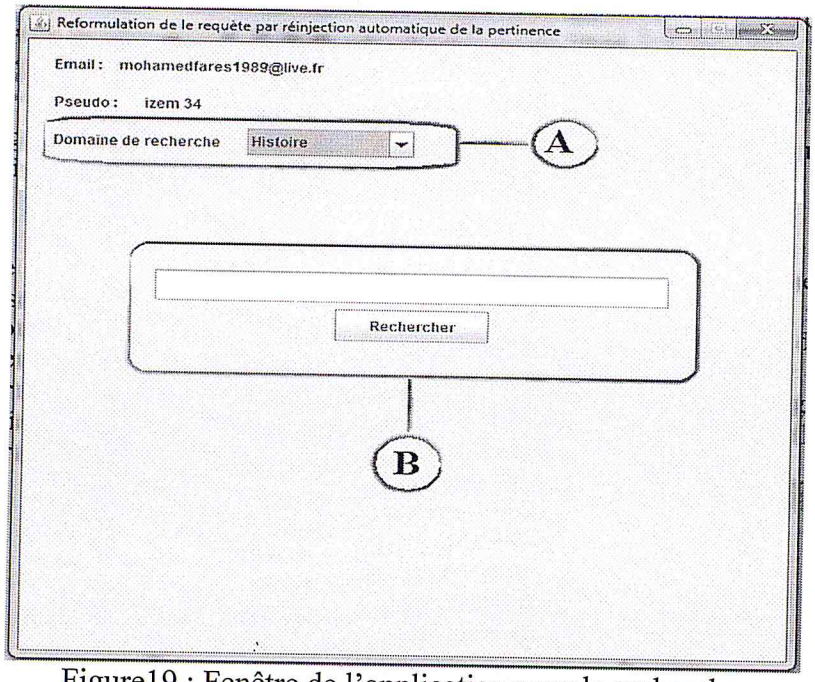

Figure19 : Fenêtre de l'application pour la recherche

5.3.1.2. Reformulation de la requête utilisateur :

Pour assurer la reformulation de la requête initiale, le module de reformulation récupère la saisie de l'utilisateur, il la compare avec le contenu de sa base de contexte et il ajoute les termes qui ont une relation avec la saisie de l'utilisateur et le domaine de recherche choisi par ce dernier. Donc la reformulation se fait après une comparaison entre les termes de la requête initiale et les termes stockés dans la base de contexte en fonction du domaine de recherche pour trouver les termes à ajouter à la requête initiale.

5.3.2. Réinjection automatique de la pertinence(RAP) :

Dans cette étape, l'application récupère le contenu des 'k' premiers documents résultants de la reformulation par le profil utilisateur. Elle extrait ensuite les évidences sans télécharger ces 'k' documents (à travers le réseau internet), et ensuite applique la formule de l'occurrence de chaque terme des documents et met tous ça dans la matrice de corrélation, puis, sélectionne la valeur max pour chaque ligne du terme de la requête et la correspondance dans les termes de colonne des documents, et ensuite concaténer les termes de la requête initiale et les termes pertinents. L'application renvoie la requête reformulée finale au moteur de recherche Google.

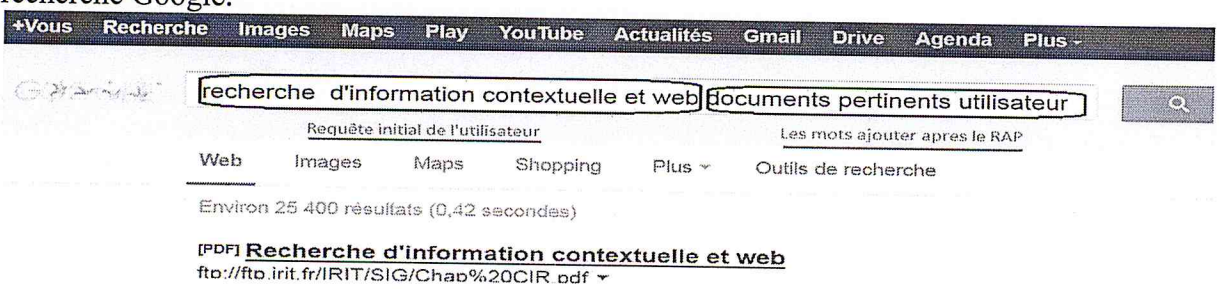

Figure20 : Fenêtre de navigateur web après la réinjection automatique de pertinence

5.4. Evaluation :

Nous nous sommes intéressés au jugement de pertinence donné par l'utilisateur pour le résultat retrouvé par le moteur de recherche Google. Ce jugement concerne l'évaluation des 50 premiers documents (qui représentent ici N) retrouvés tel que à chaque niveau de pertinence, une note de 1 ou 0 a été attribuée par chaque utilisateur : 0 correspond à un document totalement inutile ou hors-thème, 1 correspond à un document répondant de façon parfaite à la question posée. Les utilisateurs ont également exprimé leurs jugements de pertinence dans le cas où k varie entre 5,7 et 10 premiers documents retournés après la reformulation par profil utilisateur, réinjection automatique de pertinence et profil utilisateur avec réinjection automatique de pertinence, et cela pour chaque cas de l'ensemble de mot pertinent ajouté à la requête initiale (le cas d'un seul mot, de deux mots et de trois mots pertinents).

N / K					50			
		Requête initiale	Les mots pertinents pouvant être ajoutés à la requête initiale	Les mots pertinents ajoutés à la requête initiale	Nombre de documents pertinents retrouvés par la requête initiale	Nombre document pertinents trouvés avec RI par profil	Nombre document pertinents trouvés avec la RI par RAP	Nombre de documents pertinents retrouvés avec RI par profil et RAP
5	Req1	virus	malveillant	malveillant	15	31	10	38
	Req2	Type de mémoire	Vive, espace	Vive	9	25	13	32
	Req3	Recherche d'information contextuelle	Utilisateur, documents, pertinents	utilisateur	14	24	14	28
7	Req1	virus	malveillant	malveillant	15	31	10	38
	Req2	Type de mémoire	Vive, espace	Vive	9	25	10	30
	Req3	Recherche d'information contextuelle	Utilisateur, documents, pertinents	utilisateur	14	24	14	24
10	Req1	virus	malveillant	malveillant	15	31	10	38
	Req2	Type de mémoire	Vive, espace	Vive	9	25	10	30
	Req3	Recherche d'information contextuelle	Utilisateur, documents, pertinents	utilisateur	14	24	14	20

Tableau 1 Evaluation de la pertinence par le jugement de l'utilisateur selon les K premiers documents sélectionnés dans la requête initiale non reformulée, reformulation par profil, reformulation par RAP, reformulation par profil et RAP avec un seul mot pertinent ajouté à la requête initiale sur un corpus de 50(N) premiers documents retrouvés par Google.

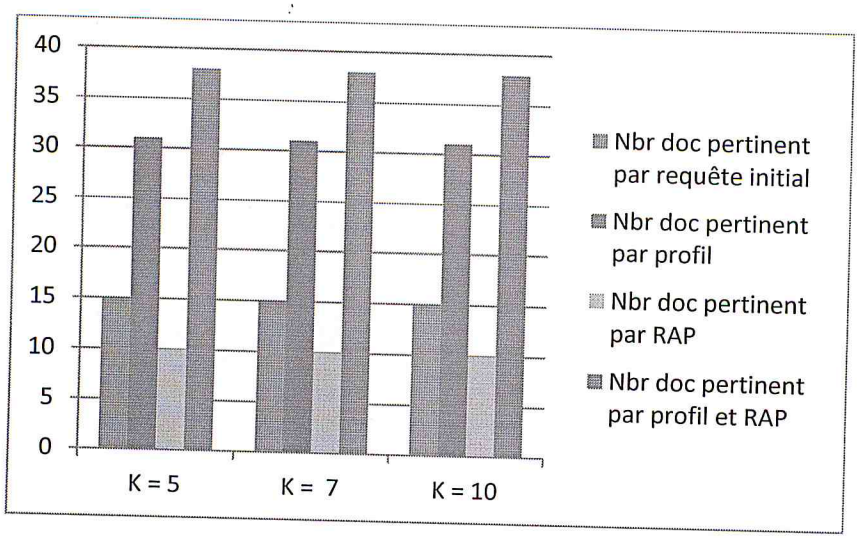

Figure21 : Graphe d'évaluation de la pertinence par le jugement de l'utilisateur selon les K premiers documents sélectionnés dans la requête initiale non reformulée, reformulation par profil, reformulation par RAP, reformulation par profil et RAP avec un seul mot pertinent ajouté à la requête initiale pour la Req1.

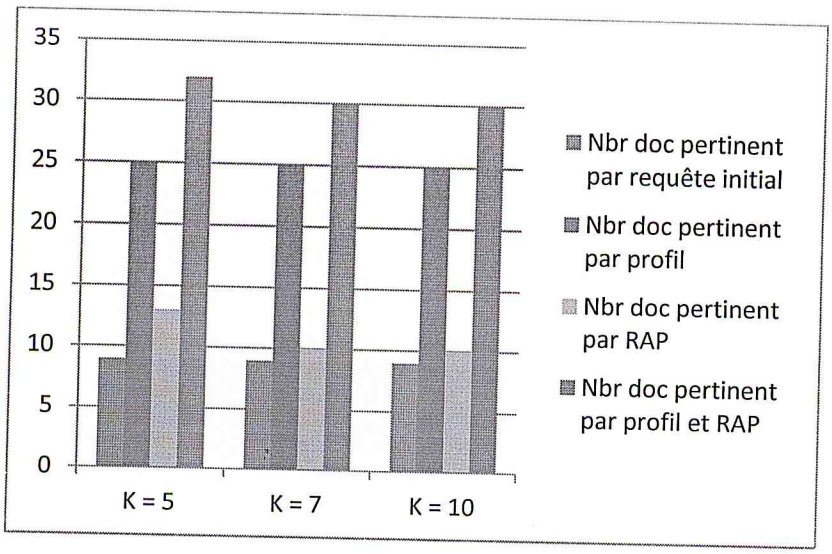

Figure22 : Graphe d'évaluation de la pertinence par le jugement de l'utilisateur selon les K premiers documents sélectionnés dans la requête initiale non reformulée, reformulation par profil, reformulation par RAP, reformulation par profil et RAP avec un seul mot pertinent ajouté à la requête initiale pour la Req2.

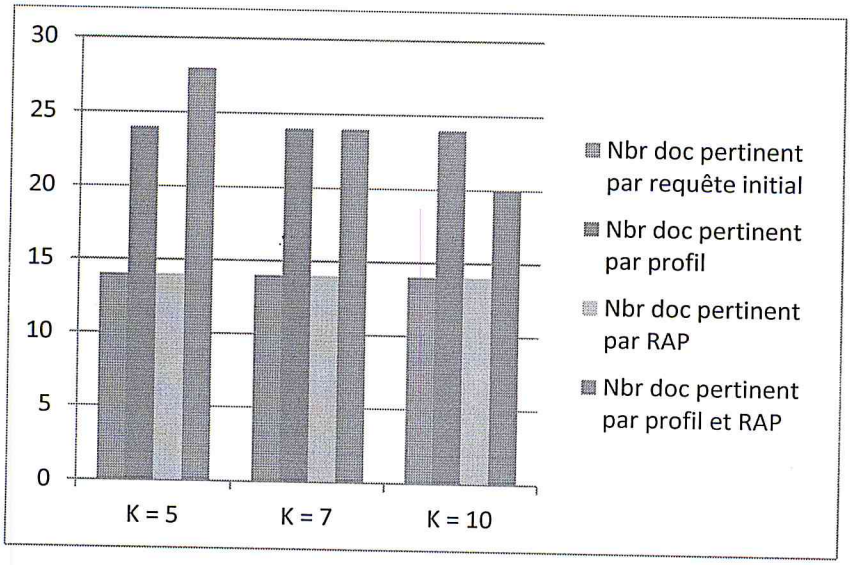

Figure23 : Graphe d'évaluation de la pertinence par le jugement de l'utilisateur selon les K premiers documents sélectionnés dans la requête initiale non reformulée, reformulation par profil, reformulation par RAP, reformulation par profil et RAP avec un seul mot pertinent ajouté à la requête initiale pour la Req3.

Nous voyons que le nombre de documents pertinents obtenus après la reformulation par profil utilisateur avec réinjection automatique de pertinence est élevé pour la requête 1 et la requête 2 contrairement à la requête 3 où ce nombre est élevé uniquement pour k=5. Le nombre de documents pertinents obtenus après la reformulation par profil utilisateur et le nombre de documents pertinents obtenus après la reformulation par réinjection automatique de pertinence restent invariables quelque soit la valeur de k dans la requête 1 et 3, excepté dans la requête 2 le cas de k=5 où il y a une augmentation au nombre de documents pertinents obtenus après la reformulation par réinjection automatique de pertinence par rapport à k=7 ou k=10,on note aussi que la reformulation par profil utilisateur donne de meilleurs résultats que la reformulation par réinjection automatique de pertinence dans ces trois requêtes.

N / K					50			
		Requête initiale	Les mots pertinents pouvant être ajoutés à la requête initiale	Les mots pertinents ajoutés à la requête initiale	Nombre de documents pertinents retrouvés par la requête initiale	Nombre document pertinents trouvés avec RI par profil	Nombre document pertinents trouvés avec la RI par RAP	Nombre de documents pertinents retrouvés avec RI par profil et RAP
5	Req1	virus	malveillant	Malveillant	15	31	10	38
	Req2	Type de mémoire	Vive, espace	Vive, espace	9	25	13	27
	Req3	Recherche d'information contextuelle	Utilisateur, documents, pertinents	Utilisateur, Documents	19	24	16	25
7	Req1	virus	malveillant	Malveillant	15	31	10	38
	Req2	Type de mémoire	Vive, espace	Vive, espace	9	25	10	28
	Req3	Recherche d'information contextuelle	Utilisateur, documents, pertinents	Utilisateur, Documents	19	24	14	19

10	**Req1**	virus	malveillant	Malveillant	**15**	**31**	**10**	**38**
	Req2	Type de mémoire	Vive, espace	Vive,espace	**9**	**25**	**10**	**25**
	Req3	Recherche d'information contextuelle	Utilisateur, documents, pertinents	Utilisateur, Documents	**19**	**24**	**16**	**20**

Tableau 2 Evaluation de la pertinence par le jugement de l'utilisateur selon les K premiers documents sélectionnés dans la requête initiale non reformulée, reformulation par profil, reformulation par RAP, reformulation par profil et RAP avec deux mots pertinents ajoutés à la requête initiale sur un corpus de 50(N) premiers documents retrouvés par Google.

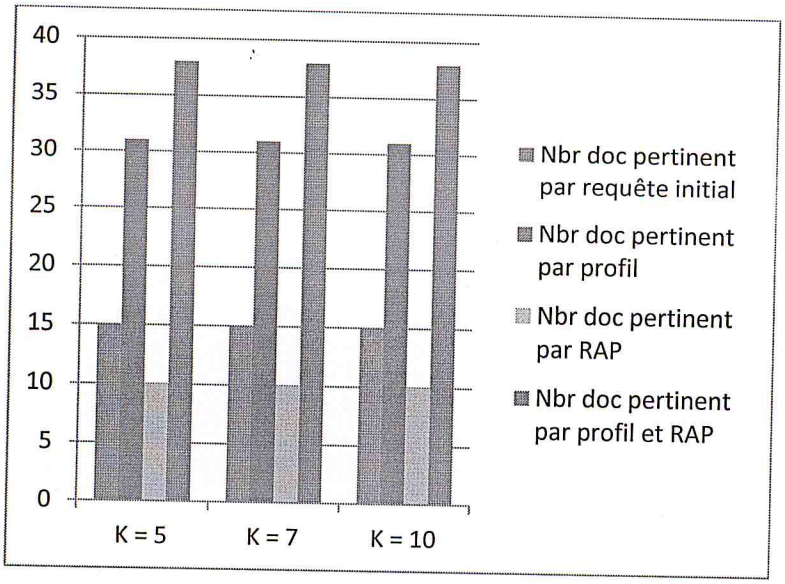

Figure24 : Graphe d'évaluation de la pertinence par le jugement de l'utilisateur selon les K premiers documents sélectionnés dans la requête initiale non reformulée, reformulation par profil, reformulation par RAP, reformulation par profil et RAP avec deux mots pertinents ajoutés à la requête initiale pour la Req1.

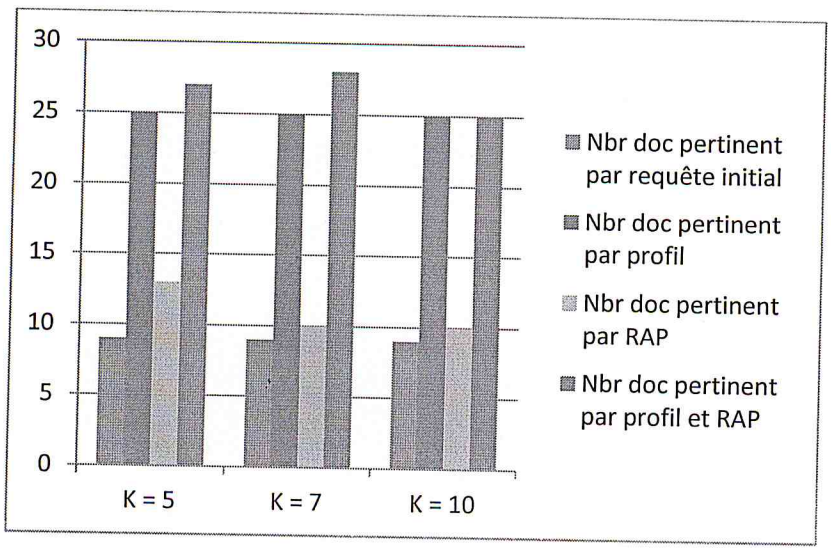

Figure25 : Graphe d'évaluation de la pertinence par le jugement de l'utilisateur selon les K premiers documents sélectionnés dans la requête initiale non reformulée, reformulation par profil, reformulation par RAP, reformulation par profil et RAP avec deux mots pertinents ajoutés à la requête initiale pour la Req2

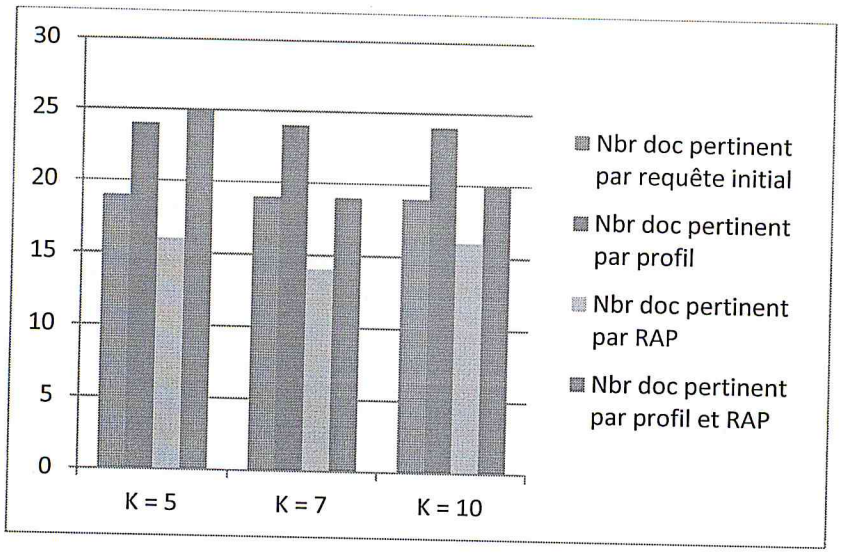

Figure26 : Graphe d'évaluation de la pertinence par le jugement de l'utilisateur selon les K premiers documents sélectionnés dans la requête initiale non reformulée, reformulation par profil, reformulation par RAP, reformulation par profil et RAP avec deux mots pertinents ajoutés à la requête initiale pour la Req3.

Nous remarquons dans ce deuxième test que pour la requête 1 les résultats sont similaires au cas précédant (le cas d'ajout d'un mot pertinent). Pour la requête 2, nous n'avons pas une augmentation sur le nombre de documents pertinents obtenus après la reformulation par profil utilisateur avec réinjection automatique de pertinence par rapport au cas précédant(le cas d'ajout d'un mot pertinent) et nous n'avons pas de changement remarquables pour le nombre de documents pertinents obtenus après la reformulation par profil utilisateur et la reformulation par réinjection automatique de pertinence. Pour la requête 3, le nombre de documents pertinents obtenus après la reformulation par profil utilisateur avec réinjection automatique de pertinence a baissé par rapport au cas précédant(le cas d'ajout d'un seul mot pertinent) dans le cas de k=5 ou k=7 et nous n'avons pas de changement remarquables pour le nombre de documents pertinents obtenus après la reformulation par profil utilisateur et la reformulation par réinjection automatique de pertinence.

N \ Y				50			
	Requête initiale	Les mots pertinents pouvant être ajoutés à la requête initiale	Les mots pertinents ajoutés à la requête initiale	Nombre de documents pertinents retrouvés par la requête initiale	Nombre document pertinents trouvés avec RI par profil	Nombre document pertinents trouvés avec la RI par RAP	Nombre de documents pertinents retrouvés avec RI par profil et RAP
Req1	Virus	malveillant	malveillant	**15**	**31**	**10**	**38**
Req2	Type de mémoire	Vive, espace	Vive,espace	**9**	**25**	**13**	**27**

5	Req3	Recherche d'information contextuelle	Utilisateur, documents, pertinents	Utilisateur, Documents, Pertinents	14	35	13	20
7	Req1	Virus	malveillant	malveillant	15	31	10	38
	Req2	Type de mémoire	Vive, espace	Vive,espace	9	25	10	28
	Req3	Recherche d'information contextuelle	Utilisateur, documents, pertinents	Utilisateur, Documents, Pertinents	14	35	18	25
10	Req1	Virus	malveillant	malveillant	15	31	10	38
	Req2	Type de mémoire	Vive, espace	Vive,espace	9	25	10	25
	Req3	Recherche d'information contextuelle	Utilisateur, documents, pertinents	Utilisateur, Documents, Pertinents	14	35	16	20

Tableau 3 Evaluation de la pertinence par le jugement de l'utilisateur selon les K premiers documents sélectionnés dans la requête initiale non reformulée, reformulation par profil, reformulation par RAP, reformulation par profil et RAP avec trois mots pertinents ajoutés à la requête initiale sur un corpus de 50(N) premiers documents retrouvés par Google.

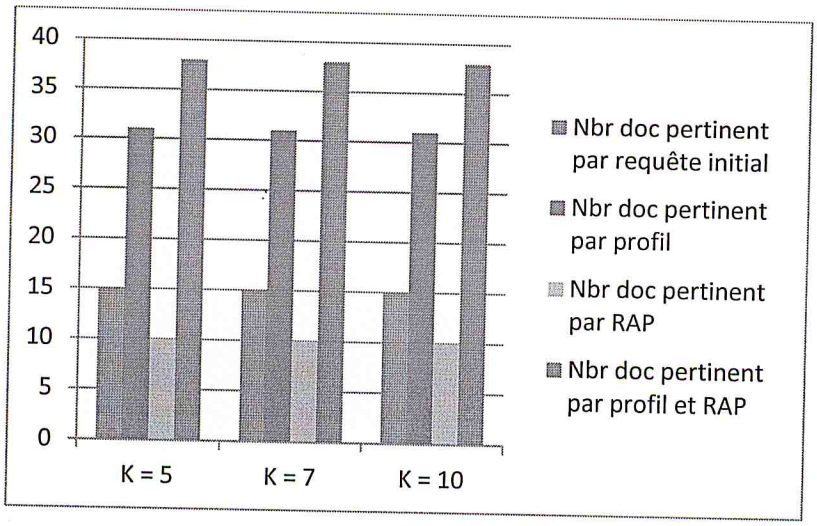

Figure27 : Graphe d'évaluation de la pertinence par le jugement de l'utilisateur selon les K premiers documents sélectionnés dans la requête initiale non reformulée, reformulation par profil, reformulation par RAP, reformulation par profil et RAP avec trois mots pertinents ajoutés à la requête initiale pour la Req1

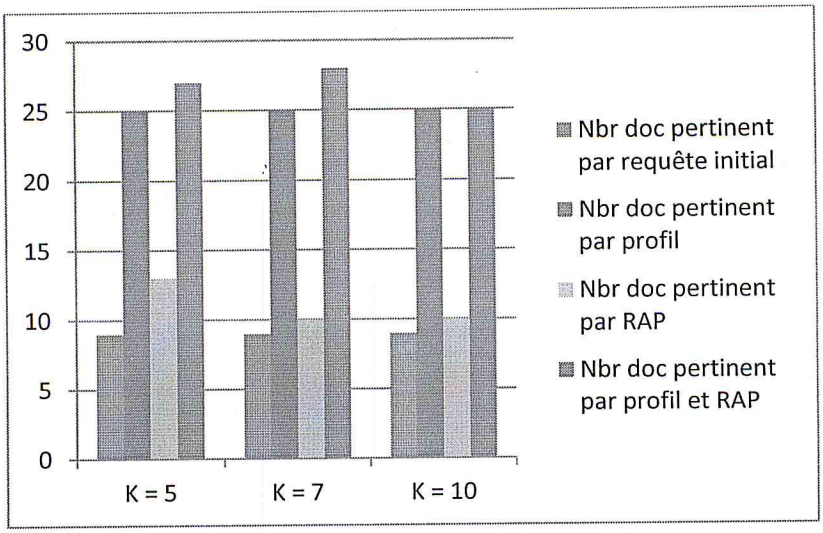

Figure28 : Graphe d'évaluation de la pertinence par le jugement de l'utilisateur selon les K premiers documents sélectionnés dans la requête initiale non reformulée, reformulation par profil, reformulation par RAP, reformulation par profil et RAP avec trois mots pertinents ajoutés à la requête initiale pour la Req2.

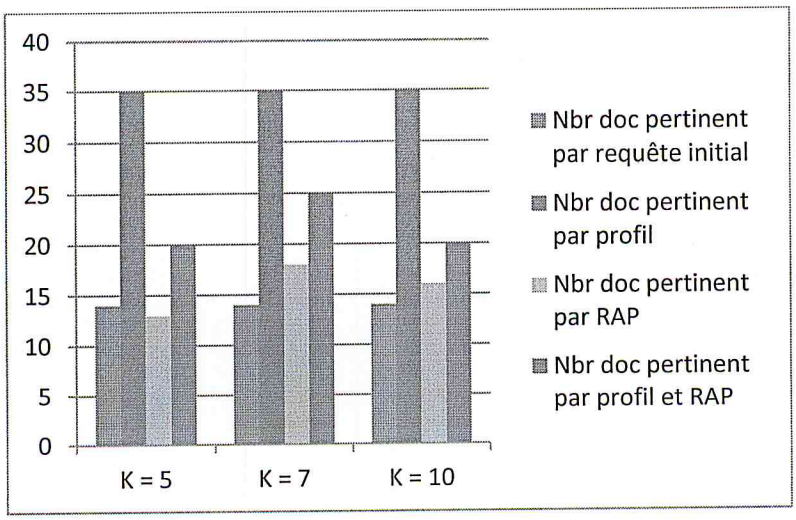

Figure29 : Graphe d'évaluation de la pertinence par le jugement de l'utilisateur selon les K premiers documents sélectionnés dans la requête initiale non reformulée, reformulation par profil, reformulation par RAP, reformulation par profil et RAP avec trois mots pertinents ajoutés à la requête initiale pour la Req3.

Pour ce troisième test, Nous remarquons que la requête 1 et la requête 2 donnent des résultats similaires aux deux cas précédents (le cas d'ajout d'un mot pertinent et le cas d'ajout de deux mots pertinents). Pour la requête 3, le nombre de documents pertinents obtenus après la reformulation par profil utilisateur avec réinjection automatique de pertinence a baissé par rapport aux cas précédents (le cas d'ajout d'un mot pertinent et d'ajout de deux mots pertinents) et nous n'avons pas de changement remarquables pour le nombre de documents pertinents obtenus après la reformulation par profil utilisateur et la reformulation par réinjection automatique de pertinence par rapport aux deux cas précédents.

Nous pouvons remarquer suivant les exemples pris que la recherche d'information par réinjection automatique de pertinence en utilisant le profil de l'utilisateur donne de meilleurs résultats que les autres approches sauf pour la requête 3. Cela s'explique du fait que ce cas conduit à une déviré de la requête.

Nous remarquons aussi que l'expansion de la requête avec plusieurs mots diminue le nombre de documents pertinents retournés ; ce qui est normal. En effet, le fait d'augmenter le nombre de mots dans une requête au delà d'un certain seuil conduit à l'effet de la dérivé de la requête. Cet effet se voit aussi dans une recherche d'information classique.

A la fin, nous pouvons conclure que notre approche donne des résultats acceptables surtout en prenant k à 5.

5.5. Conclusion

Dans ce chapitre, nous avons abordé le mécanisme en général de notre système avec présentation de l'aspect graphique de l'application visible par l'utilisateur, au cours de l'explication de ce fonctionnement, nous avons montré le rôle joué par le profil utilisateur dans le processus général de reformulation de requête sans oublier la réinjection de pertinence. Puis, nous avons fait un tour d'horizon vers les différents outils et langages utilisés lors de l'implémentation du projet. A la fin, nous avons présenté une étude d'évaluation dans un contexte donné où nous avons résumé les résultats du test dans les tableaux représentatifs associés suivis par les graphes reflétant ces derniers.

Conclusion générale

Conclusion générale :

Notre travail se situe dans le cadre de la réinjection automatique de la pertinence dans la recherche d'information contextuelle. Plus précisément on a travaillé par le profil utilisateur (la recherche d'information personnalisée). Nous avons alors proposé un mécanisme de réinjection automatique de pertinence utilisant le profil utilisateur.

La réinjection automatique de pertinence se fait par ajout des termes extraits d'un échantillon de documents. Mais la limite majeur de cette méthode est lorsque cet échantillon ne contient pas des documents pertinents. Pour résoudre ce problème et pour garder l'idée générale de la réinjection automatique de pertinence, nous avons introduit le profil utilisateur pour obtenir un échantillon pertinent mais sans l'intervention directe de l'utilisateur. Nous avons choisis la modélisation multidimensionnelle pour représenter le profil utilisateur, ensuite nous avons exploité ce profil pour trouver les « k » premiers documents pertinents, l'exploitation a été faite par la reformulation de requêtes. Donc en résumé, nous avons fait une première reformulation par profil utilisateur pour trouver les documents pertinents et ensuite nous faisions l'extraction des termes à ajouter à la requête initiale à partir de ces derniers avec une méthode de pondération qu'on a choisi qui a pour but final la construction d'une nouvelle requête en combinant la requête initiale avec les informations extraites.

Il est clair que la recherche d'informations contextuelle est un domaine de recherche très vaste et couvre plusieurs aspects qu'il serait intéressant d'approfondir, tels que :

- La prise en compte d'autres types de documents et pas seulement les documents textuels ;
- La prise en compte d'autres langues, dans notre système nous avons utilisé des techniques permettant la manipulation des documents en français seulement ;
- La prise en compte des mots composés dans toutes les étapes d'indexation, y compris leurs reconnaissances, leurs normalisations et leurs pondérations ;
- Améliorer la représentation du profil utilisateur, pour permettre sa mise à jour en prenant en considération son feedback avec le système ;
- Utilisation d'autres types de reformulation basés sur le profil : filtrage social et hybride ou intégration des thesaurus et des ontologies. Pouvoir générer automatiquement des sites portails personnalisés et adaptés à chaque profil ;
- Utilisation des règles d'associations comme une approche pour choisir les termes ajoutés à la requête initiale.

Glossaire

Web : [39]
Le World Wide Web, communément appelé le Web, parfois la Toile, littéralement la « toile (d'araignée) mondiale », est un système hypertexte public fonctionnant sur Internet et qui permet de consulter, avec un navigateur, des pages mises en ligne dans des sites. L'image de la toile vient des hyperliens qui lient les pages Web entre elles.
Le Web n'est qu'une des applications d'Internet, avec le courrier électronique, la messagerie instantanée, Usenet, etc. Le Web a été inventé plusieurs années après Internet, mais c'est le Web qui a rendu les médias grand public attentifs à Internet. Depuis, le Web est fréquemment confondu avec Internet ; en particulier, le mot Toile est souvent utilisé de manière très ambiguë.
Le World Wide Web est et a été désigné par de nombreux noms et abréviations synonymes : WorldWideWeb, World Wide Web, World-wide Web, Web, WWW, W3, Toile d'araignée mondiale, Toile mondiale, Toile.

URL : [40]
URL signifie Uniform Ressource Locator Il s'agit pour résumer de l'adresse unique qui permettra d'ouvrir votre page, ou un fichier particulier. Dans le domaine du référencement et de la veille, l'url est un facteur essentiel, puisque c'est elle qui est enregistrée et présentée aux internautes. Cette adresse débute par le protocole (http://), qui définit el langage utilisé sur le réseau. Le plus utilisé est sans nul doute le protocole http, qui vise à échanger des fichiers html. Il existe d'autres protocoles, comme FTP, MAILTO, HTTPS ...

Navigateur : [41]
Un navigateur web est un logiciel informatique qui permet d'utiliser le web. Pour être plus précis, ce type de logiciel permet de consulter le *World Wide Web* (WWW). L'utilisation la plus répandue de ces logiciels étant de visualiser les pages web et d'utiliser les liens hypertextes dans le but d'aller de pages en pages. Il s'agit d'un logiciel possédant une interface graphique composée de boutons de navigation, d'une barre d'adresse, d'une barre d'état (généralement en bas de fenêtre) et dont la majeure partie de la surface sert à afficher les pages web.

Google : [42]
Google est le moteur de recherche le plus populaire au monde. Il a commencé comme un projet de recherche en 1996 par Larry Page et Sergey Brin, qui étaient deux Ph.D. étudiants de l'Université Stanford. Ils ont développé un algorithme du moteur de recherche qui classe les pages Web non seulement par le contenu et les mots clés, mais par combinaison d'autres pages Web liées à chaque page. Cette stratégie a donné des résultats plus utiles que d'autres moteurs de recherche, et a conduit à une augmentation rapide de la recherche Web parts de marché de Google. L'algorithme de classement de Google a été nommé plus tard "PageRank" et a été breveté en Septembre 2001. En peu de temps, Google est devenu le leader des moteurs de recherche dans le monde.

Internet : [43]
Internet est le réseau informatique mondial qui rend accessibles au public des services variés comme le courrier électronique, la messagerie instantanée et le World Wide Web, en utilisant le protocole de communication IP (*internet protocol*). Son architecture technique qui repose sur une hiérarchie de réseaux lui vaut le surnom de réseau des réseaux.

HTML : [43]
Hyper Text Mark-Up Language. Langage décrivant un contenu hypertextuel : il permet d'indiquer dans une page où placer le texte, les images, les vidéos, les liens vers d'autres ressources, etc. Proposé en 1989 Par Tim Berners-Lee, ce format, qui a donné naissance au World Wide Web, connaît aujourd'hui plusieurs versions.

Ontologie : [44]
La définition des ontologies est héritée d'une tradition philosophique qui s'intéresse à la science de l'Etre. Aujourd'hui, elle signifie la « science des étants » c'est-à-dire l'ensemble des objets reconnus comme existants dans un domaine.
L'ontologie est utilisée, depuis plusieurs années, dans l'Ingénierie des Connaissances (IC) et l'Intelligence Artificielle (IA) pour structurer les concepts d'un domaine. Les concepts sont rassemblés et ces derniers sont considérés comme des briques élémentaires permettant d'exprimer les connaissances du domaine qu'il recouvre.
Les ontologies sont utiles pour partager des connaissances, créer un consensus, construire des systèmes à base de connaissances. De nombreux projets d'ontologies sont en œuvre comme celle du Web sémantique. Le problème fondamental est de respecter la diversité des langages et des représentations du monde, tout en permettant les échanges d'informations.

Bibliographie

[1] RESSAD-BOUIDGHAGHEN Ourdia « Accès contextuel à l'information dans un environnement mobile : approche basée sur l'utilisation d'un profil situationnel de l'utilisateur et d'un profil de localisation des requêtes. », Thèse de doctorat, Université de Paul Sabatier Toulouse, 2011.

[2] Mariam DAOUD « Accès personnalisé à l'information : approche basée sur l'utilisation d'un profil utilisateur sémantique dérivé d'une ontologie de domaines à travers l'historique des sessions de recherche. », Thèse de doctorat, Université de Paul Sabatier Toulouse, 2009.

[3] ABDELKRIM BOURAMOUL « Recherche d'information contextuelle et sémantique sur le web. », Thèse de doctorat, Université de Constantine, 2011.

[4] Ba-Duy DINH « Accès à l'information biomédicale : vers une approche d'indexation et de recherche d'information conceptuelle basée sur la fusion de ressources termino-ontologiques», Thèse de doctorat, Université de Paul Sabatier Toulouse, 2012.

[5] WAHIBA NESRINE ZEMIRLI « Modèle d'accès personnalisé à l'information basé sur les Diagrammes d'Influence intégrant un profil utilisateur évolutif», Thèse de doctorat, Université de Paul Sabatier Toulouse, 2008.

[6] Lobna HLAOUA « Reformulation de Requêtes par Réinjection de Pertinence dans les Documents Semi-Structurés », Thèse de doctorat, Université de Paul Sabatier Toulouse, 2007.

[7] WAHIBA NESRINE ZEMIRLI « Vers le développement d'un système de recherche d'information personnalisé intégrant le profil utilisateur», Thèse de doctorat, Université de Paul Sabatier Toulouse, 2004.

[8] Nawel Nassr « Croisement de langues en recherche d'information : traduction et désambiguïsation de requêtes », Thèse de doctorat, Université de Paul Sabatier Toulouse, 2002.

[9] Fatiha BOUBEKEUR-AMIROUCHE « Contribution à la définition de modèles de recherche d'information flexibles basés sur les CP-Nets », Thèse de doctorat, Université de Paul Sabatier Toulouse, 2008.

[10] Fabien Picarougne « Recherche d'information sur Internet par algorithmes évolutionnaires », Thèse de doctorat, Université de François Rabelais Tours, 2004.

[11] Lynda TAMINE-LECHANI « De la recherche d'information orientée système vers la recherche d'information orientée contexte : Verrous, contributions et perspectives », Thèse de doctorat, Université de Paul Sabatier Toulouse, 2008.

[12] Hicham CHEBILI «Agrégation des résultats dans la recherche d'information Semi-Structurée », Mémoire de Magister, Ecole supérieure d'informatique (E.S.I) Oued-Smar Alger, 2011.

[13] Lynda Lechani Tamine, Mohand Boughanem « Accès personnalisé à l'information : Approches et Techniques », Rapport interne, Institut de recherche en Informatique de Toulouse, 2005.

[14] Anis Benammar « Proposition à l'intégration des profils dans le processus de recherche d'information », Article, Institut de recherche en Informatique de Toulouse, 2010.

[15] Hugues Bouchard et Jian-Yun Nie « Modèles de langue appliqués à la recherche d'information contextuelle », Article, Département de recherche opérationnelle (DIRO), Université de Montréal, 2006

[16] Lynda Tamine-Lechani, Nesrine Zemirli, Wahiba Bahsoun « Approche statistique pour la définition du profil d'un utilisateur de système de recherche d'information », Article, Institut de Recherche en Informatique de Toulouse, 2008.

[17] Victor Odumuyiwa « De la recherche sociale d'information à la recherche collaborative d'information », Article, Laboratoire de recherche en informatique et ses applications de Nancy, 2009.

[18] Victor Odumuyiwa « La gestion de la recherche collaborative d'information dans le cadre du processus d'intelligence économique » Thèse de doctorat, Université de Nancy, 2011.

[19] Madalina Mitran « Recherche d'information sociale : exploitation du social bookmarking pour enrichir l'accès à l'information », Mémoire de master, université Paul Sabatier Toulouse, 2010.

[20] Robin VIVIAN, Jérôme DINET « un système collaboratif de recherche d'information centré utilisateur », Article, université Paul Verlaine de Metz, 2008.

[21] Lynda TAMINE-LECHANI, Sylvie CALABRETTO « Recherche d'information contextuelle et web », Article, université Paul Sabatier Toulouse, 2008.

[22] Med El Amine Abderrahim, Med Alaeddine Abderrahim « Réinjection automatique de la pertinence pour la recherche d'informations dans les textes Arabes », Article, Université de Tlemcen Algérie, 2012.

[23] Lynda Tamine « optimisation de requêtes dans un système de recherche d'information approche basée sur l'exploitation de techniques avenacées de l'algorithmique », Thèse de doctorat, Université Paul Sabatier de Toulouse, 2000.

[24] S. Gauch et J. Wang « analyse de corpus textuels de TREC 5 requêtes Expansion », Article dans les procédures de la 5e Conférence de recherche d'information, USA, 1996

[25] http://ipeti.forumpro.fr/t21-definition-de-langage-java-java-script

[26] J. Xu, W.B Croft « Extension de requête utilisant l'analyse de documents locale et globale », Article, Conférence de recherche et de développement, Université de Zurich Germany, 1996.

[27] Hassan CHOUAIB « Reformulation de requêtes dans un modèle de réseau possibiliste pour la recherche d'information », Mémoire DEA d'Informatique, Université Paul Sabatier, 2006.

[28] J.J. Rocchio « Pertinence des commentaires dans la recherche d'information. Dans La récupération système SMART » expériences dans le traitement automatique de documents, 1971.

[29] D. Harman « Vers l'expansion de requête interactive », En 11e Conférence internationale annuelle ACM SIGIR sur la recherche et développement en Recherche d'information, 1988

[30] S.robertson, K.Sparck Jones « La pondération de la pertinence des termes de recherche» Article, Journal de la Société américaine pour les sciences de l'information, 1976.

[31] H. Jing et E. Tzoukermann « recherche d'information sur le contexte Distance et morphologie », Article, Conférence sur la recherche et le développement dans l'information Récupération, USA, 1999.

[32] MARTINET Jue « Un modèle vectoriel relationnel de recherche d'information adapté aux images » Thèse de doctorat, Université Joseph Fourier , Grenoble I. 2004.

[33] ROUSSEY Cyte « Une méthode d'indexation sémantique adaptée aux corpus multilingue », thèse de doctorat, Institut National des Sciences Appliquées(INSA) de Lyon, 2001.

[34] Samiha El Hamali « Système de filtrage d'informations pour la gestion de crises et de catastrophes », diplôme de Magister, Ecole Nationale Supérieure d'Informatique ESI, Alger, 2012.

[35] Baziz Mohamed « Indexation conceptuelle guidée par ontologie pour la recherche d'information », thèse de doctorat, IRIT, Toulouse, France. 2005.

[36] http://www.techno-science.net/?onglet=glossaire&definition=517

[37] http://www.dicofr.com/cgi-bin/n.pl/dicofr/definition/20020117173109

[38] http://www.craym.eu/tutoriels/developpement/site_local_avec_wamp.html

[39] http://www.adproxima.fr/glossaire-5-www.html

[40] http://www.docpc-informatique.com/rb42-l30-definition-url.html

[41] http://glossaire.infowebmaster.fr/navigateur-web/

[42] http://www.techterms.com/definition/google

[43] http://www.ritimo.nursit.com

[44] http://www.technolangue.net/imprimer.php3?id_article=280