الجمهوريه الجزائريه الديمقراطيه الشعبيه

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Yahia Fares-Médéa

Faculté des Sciences

Département Mathématique et Informatique



Projet de fin d'étude

Présenté pour l'obtention du Diplôme de

Master en Informatique

Option : Ingénierie des Systèmes et Technologies du Web

THEME

Reformulation sémantique et personnalisée de la requête d'utilisateur pour améliorer la pertinence des résultats de recherche d'information

Proposé et Encadré par :

-Mr.Rafa Tahar

Préparé et présenté par :

-Cheriet Asrar

-Amirat Sarah

2019-2020

Résumé :

Le but principal des systèmes de recherche d'information (SRI) est de retrouver plus des documents pertinents pour une requête d'utilisateur, à partir d'une base de documents volumineuse.

Tandis que les documents retourné par les systèmes de recherches d'information classique ne sont pas forcément pertinents par rapport à l'utilisateur même si le système les juge comme pertinents.et pour résoudre le problème de l'écart entre la pertinence système et la pertinence utilisateur les approches de recherche d'informations actuelle s'orientent vers la recherche d'information personnalisée qui visent à l'intégration de l'utilisateur dans le processus de recherche via la modélisations des utilisateurs et leurs préférences .

L'objective de la RI personnalisée est de faciliter l'expression du besoin utilisateur et de rechercher des informations sur un sujet en écartant l'information non pertinente et donc réduire l'espace de recherche et mieux répondre au besoin en information.

Dans ce travail on va présenter une architecture et une évaluation d'un système de recherche d'information personnalisée en intégrant la sémantique dans le profile utilisateur.

La représentation de profile utilisateur est fait selon une matrice qui contient les termes extrait de l'historique de recherche de l'utilisateur, les concepts extrait de l'ontologie wordnet lors de la projection des termes de profil sur l'ontologie et la similarité sémantique entre les termes et les concepts.

Abstract :
The main purpose of information retrieval systems (IRS) is to find more documents relevant to a user query,from a large database.

While the documents returned by traditional information retrieval systems are not necessarily relevant for the user even if the system considers them as relevant. to solve the problem of the gap between system relevance and user relevance, current information retrieval approaches are geared towards personalized information retrieval, which means integrating the user into the search process through the modeling of users and their preferences.

The objective of personalized information system is to facilitate the expression of user needs by removing irrelevant information therefore reduce the search space and satisfy the users by finding more relevant information.

In this work we will present architecture and an evaluation of a personalized information retrieval system by integrating the semantic into the user profile.

The representation of the user profile is made according to a matrix which contains the relevant terms extracted from the user's search history, theconcepts extracted from the wordnet ontology when projecting profile terms on the ontology and semantic similarity between terms and concepts.

ملخص

الغرض الرئيسي من أنظمة استرجاع المعلومات (IRS) هو العثور على المزيد من المستندات ذات الصلة باستعلام المستخدم ، من قاعدة بيانات كبيرة.

في حين أن المستندات التي يتم إرجاعها بواسطة أنظمة استرجاع المعلومات التقليدية ليست بالضرورة ذات اهميه بالنسبه للمستخدم حتى لو اعتبرها النظام ذات اهميه . ولحل مشكلة الفجوة بين نظام استرجاع المعلومات و المستخدم وأهميته يتم توجيه مناهج بحث المعلومات الحالية للمستخدم نحو البحث عن المعلومات الشخصية مما يعني دمج المستخدم في عملية البحث عبر نمذجة المستخدمين وتفضيلاتهم.

الهدف من نظام استرجاع المعلومات الشخصي هو تسهيل التعبير عن احتياجات المستخدم والبحث عن معلومات حول موضوع ما من خلال التخلص من المعلومات التي ليس لها صله بالموضوع وبالتالي تقليل مساحة البحث واسترداد المعلومات المطلوبة بشكل أفضل.

في هذا العمل ، سنقدم بنية وتقييمًا لنظام استرجاع المعلومات المخصص من خلال دمج الدلالات في ملف تعريف المستخدم.يتم تمثيل ملف تعريف المستخدم وفقًا لمصفوفة تحتوي على المصطلحات المستخرجة من سجل البحث الخاص بالمستخدم ،والمفاهيم المستخرجة من الأنطولوجيا Wordnet عند عرض مصطلحات الملف الشخصي على الأنطولوجياباالاضافه الى التشابه الدلالي بين المصطلحات والمفاهيم.

Table des matières

Introduction générale 8

Chapitre 1 : la recherche d'information classique et sémantique 11

1 INTRODUCTION : 11

2. Les fondements de la recherche d'information (RI) 12

2.1 Concepts de base de la recherche d'information : 12

2.2 Système de recherche d'information(SRI) : 15

3. De la RI classique à la RI adaptative : 30

3.1 La RI adaptative : 31

3.2 Limitations de la RI adaptative : 33

4 La recherche d'information sémantique : 33

4.1 Définition : 33

4.2 Indexation sémantique : 33

4.3 Les ressources linguistiques externes : 35

4.4 Les approches d'indexation sémantique basée sur la désambiguïsation : 36

4.5 Les approches de désambiguïsation des sens des mots (WSD) : 37

5 CONCLUSION : 38

Chapitre2 : recherche d'information personnalisée 40

1. INTRODUCTION : 40

2. Notions de base : 41

2.1 Contexte de recherche : 41

2.2 Pertinence contextuelle : 41

2.3 Profil utilisateur : 41

3 Les services de personnalisation : 43

3.1 Le filtrage des résultats : 43

3.2 Le re-ordonnancement des résultats de la requête : 44

3.3 Reformulation de la requête : 45

3.4 La recommandation : 45

4 System de recherche d'information personnalisée : 46

4.1 Définition : 46

4.2 Architecture générale (SRIP) : 46

4.3 Les prototypes des systèmes d'accès personnalisés : 55

4.4 Approches de modélisation du profil utilisateur : 56

5 Évaluation d'accès personnalisée àl'information : 60
5.1 Problèmes de l'évaluation :....61
5.2 Recommandations : 61
6 Contribution recherche sémantique et personnalisée :....62
6.1 Représentation sémantique du profil utilisateur :....62
6.2 Pondération sémantique des termes 64
7 CONCLUSION :....65
Chapitre 3 : Reformulation de la requête....67
1. INTRODUCTION : 67
2. La reformulation classique : 68
2.1 La reformulation manuelle :....68
2.1 La reformulation semi-automatique :....69
3. La reformulation sémantique : 69
3.1 Les Ressources sémantiques : 70
3.2 Quelques approches de reformulation sémantique : 74
4. La reformulation personnalisée : 75
4.1 Exploiter le profil d'utilisateur pour la reformulation :....76
4.2 Quelques approches de reformulation personnalisée :....76
5. Les defaults des approches de la reformulation des requêtes : 78
6. CONCLUSION :....79
Chapitre 4 :reformulation sémantique et personnalisée de la requête d'utilisateur....81
1. INTRODUCTION : 81
2. L'ontologie utilisé : 81
2.1 WordNet : 82
3. Le profil utilisateur proposé : 84
3.1 Partie statique : 85
3.2 Partie dynamique : 86
3.3 Partie sémantique :....87
4. Processus de recherche d'information personnalisée : 90
4.1 L'enrichissement de la requête :....91
4.2 Mise à jour de profil :....93
5. Evaluation de l'approche :....94
5.1 Stratégie d'évaluation :....94

5.2 L'expérimentation :.....95

5.3 Résultats et discussion :.....99

6. CONCLUSION :.....104

Conclusion générale.....105

REFERENCES :.....106

Table des figures:

Figure 1.1:Système de Recherche d'Information……...15

Figure1.2:Processus de recherche d'information……..…16

Figure 1.3:les règles de correspondances du model booléen……20

Figure 1.4:les mesures de similarités de modèle vectoriel de base……22

Figure 1.5:Partition de la collection pour une requête……...27

Figure 1.6:graphe d'une courbe rappel –précision……..…28

Figure 2.7:Représentation d'un système de filtrage……..……43

Figure *2.1*: Architecture générale d'un système de recommandation……44

Figure 2.*3* : Architecture générale d'un SRIP……46

Figure 3.1: Exemple d'une taxonomie des animaux……. 70

Figure 3.2:Le mot « book » dans le thésaurus anglais WordNet ……..…72

Figure 3.3:Un extrait de l'ontologie……..……73

Figure 4.1 : les concepts de Wordnet correspondants au mot mouse……81

Figure 4.2 :Sous hiérarchie de WordNet correspondant au concept "dog"……...82

Figure 4.3: présentation de profil Utilisateur……84

Figure 4.4: présentation de la partie statique……85

Figure 4.5: présentation de la partie dynamique……85

Figure 4.6: présentation de la partie sémantique……..86

Figure 4.7 : représentation des résultats de la requête « pièce classique » avant la reformulation……*100*

Figure 4.8 : représentation des résultats de la requête « piece classique » après la reformulation en utilisant le 1er scénario……..101

Figure 4.9 : représentation des résultats de la requête « piece classique » après la reformulation en utilisant le 2ème scénario……..102

Introduction générale

La recherche d'information (RI), est une branche del'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.

Aujourd'hui, interroger une ressource d'information par une requête composée de quelques mots clés n'est plus une pratique réservée aux spécialistes de certains domaines. Cette pratique est devenue un réflexe qui concerne tout individu ayant un accès à Internet. De plus, le développement rapide des outils de communication et des technologies de stockage, ainsi que la simplicité technique de publier une information sur Internet, ont permis la production d'une masse géante d'informations sur le Web pour laquelle on ne dispose souvent pas de garantie de qualité ou de fiabilité. Pour cela, mettre en correspondance un besoin d'information et un document pertinent est encore plus complexe aujourd'hui même si cela a toujours été la problématique de base de la recherche d'information.

L'objectif de la RI personnalisée est de mieux répondre aux besoins en information de l'utilisateur tout en intégrant le contexte de recherche dans la chaîne d'accès à l'information. De même la RI sémantique est généralement guidée par les ontologies. L'utilisation d'ontologie en RI a pour finalité de spécifier des connaissances qui seront interprétables à la fois par l'homme et par la machine. Le contexte et la sémantique constituent les deux notions principales représentant les mécanismes que nous utilisons comme support à la modélisation de l'utilisateur et ses intérêts.

Dans cemémoire, nous nous intéressons à la recherche d'information dans un contexte Web, donc on a proposé une combinaison de deux approche de recherche d'information (personnalisée et sémantique) .cette combinaison s'appuie sur un profile utilisateur sous forme sémantique ou les termes pertinentes sont extraits à partir des documentes pertinentes de l'historique de recherche pour former le centre d'intérêt de l'utilisateur, ainsi en utilisant des concepts extrait de l'ontologie *WordNet*afin de désambiguïser sémantiquement les termes pertinents. Le but principal de cette approche est l'exploitation des préférences de l'utilisateur extraitesde son profil dansle processus de recherche pour recevoir comme résultat une liste ordonnée de documents d'une manière décroissante selon leur pertinence évaluée par le système.

Le document est organisé comme suit :

Dans le premier chapitre nous avons entamé Les fondements de la recherche d'information classique et sémantique, et pour le deuxième chapitre on a présenté la recherche d'information personnalisée et donner quelques représentations sémantiques de profil dans la recherche personnalisée. Le troisième chapitre est consacré pour la reformulation (classique, personnalisée est sémantique) des requêtes utilisateur. Dans le quatrième chapitre est présentée notre contribution .et nous finissons par une conclusion générale.

Chapitre 1 :
La recherche d'information classique et sémantique

Chapitre 1 : la recherche d'information classique et sémantique

1 INTRODUCTION :

La Recherche d'Information (RI) est un domaine qui existe depuis plusieurs décennies. Les premières recherches de la RI ont été initiés par Salton G. qui a établi cette définition : « la recherche d'information est un domaine qui a pour objectif, la représentation, l'analyse, l'organisation, le stockage et l'accès à l'information »

L'opérationnalisation de la RI est réalisée par des outils informatiques appelés Systèmes de Recherche d'Information (SRI) ou un engin de recherche. Ce dernier permet de retourner les documents, Parmi ces documents retournés certains sont pertinents, mais noyés parmi beaucoup d'autres documents non pertinents à une requête d'utilisateur, à partir d'une base de documents volumineuse.

Dans ce chapitre, nous allons entamer Les fondements de la recherche d'information (RI), et dans cette partie nous allons donner un aperçu sur les concepts de base de la recherche d'information puis aborder le Système de recherche d'information(SRI) avec ces composantes qui sont détaillées, on débutant avec le processus de recherche puis Les différents modèles de recherche d'information, ensuite Evaluation des performances des systèmes de RI, et nous présentons la RI adaptative et on finissants avec sa limitations et en fin en s'intéresse a la méthode sémantique dans la recherche classique en utilisant des approches d'indexation sémantique

2. Les fondements de la recherche d'information (RI)

L'accès aux informations pertinentes pour un utilisateur parmi celles stockées dans une base documentaire avec une gigantesque masse de données devient alors une tâche pénible. Afin de traiter ce problème, une discipline toute entière est née. Cette discipline est appelée Recherche d'**i**nformation (RI). [Ameni YENGUI,2016]

Avec l'augmentation rapide du volume documentaire stocké sous format numérique, il est devenu très difficile de trouver une information ou un document qui répond à un besoin utilisateur. La recherche d'Information (RI) est une discipline relativement ancienne qui s'intéresse à répondre à la problématique : comment retrouver un document parmi une collection de documents qui satisfait un besoin utilisateur. [Rami HARRATHI,2010]

La recherche d'information (RI) est un champ qui est destiner à l'apprentissage ,l'organisation de stockage, et au développement des méthodes qui permettent de trouver l'information nécessité par l'utilisateur en informations pertinentes.

2.1 Concepts de base de la recherche d'information :

Le but de la recherche d'information est d'utiliser des techniques et des moyens afin de trouver les informations pertinentes contenues dans une quantité massive des documents, en réponse aux besoins en information d'un utilisateur représentés à l'aide d'une requête formulée par ce dernier.

Nous citons ci-dessous la définition de la RI donnée par [C.VAN Rijsbergen,1986] :

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

1.1.1 Document :

Un document peut être un texte, un morceau de texte, une page web, une image, une vidéo, etc. On peut appeler document toute unité qui peut constituer une réponse à un besoin informationnel de l'utilisateur. [NesrinZemirli,2009]

Le document est l'élément central de SRI. C'est un objet complexe sans cesse en évolution car il est lié aux développements des technologies de la communication. Un document peut être un texte, un morceau de texte, une page web, une image, une vidéo, etc. On peut appeler document toute unité qui peut constituer une réponse à un besoin informationnel de l'utilisateur. [Mahiddine Mehanna,2007]

Le document constitue l'information élémentaire d'une collection de documents

1.1.2 Collection de documents :

La collection de documents (ou fond documentaire, corpus) constitue l'ensemble des informations (des documents) exploitables et accessibles. [Ourdia,2011]

Dans le cas général et pour un souci d'optimalité, la base constitue des représentations simplifiées mais suffisantes pour ces documents. [BOURAMOUL,2011]

1.1.3 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. [Mariam Daoud ,2008]

1.1.4 Besoin d'information :

La notion de besoin en information en recherche d'informations est souvent assimilée au besoin de l'utilisateur. Trois types de besoin utilisateur ont été définis par [Ingwersen, 92] :

- Besoin vérificatif :L'utilisateur cherche à vérifier le texte avec les données connues qu'il possède déjà. Il recherche donc une donnée particulière, et sait même souvent comment y accéder. La recherche d'un article sur Internet à partir d'une adresse connue serait un exemple d'un tel besoin. Un autre exemple serait de chercher la date de publication d'un ouvrage dont la référence est connue. Un besoin de type vérificatif est dit stable, c'est à dire qu'il ne change pas au cours de la recherche.
- Besoin thématique connu :L'utilisateur cherche à clarifier, à revoir ou à trouver de nouvelles informations dans un sujet et un domaine connus. Un besoin de ce type peut être stable ou variable ; il est très possible en effet que le besoin de l'utilisateur s'affine au cours de la recherche. Le besoin peut aussi s'exprimer de façon incomplète, c'est-à-dire que l'utilisateur n'énonce pas nécessairement tout ce qu'il sait dans sa requête mais seulement un sous-ensemble. C'est ce qu'on appelle dans la littérature le label.
    - Besoin thématique inconnu :Cette fois, l'utilisateur cherche de nouveaux concepts ou de nouvelles relations en dehors des sujets ou des domaines qui lui sont familiers. Le besoin est intrinsèquement variable et est toujours exprimé de façon incomplète.

1.1.5 Pertinence :

La pertinence est une notion fondamentale dans le domaine de la RI.et d'après [M. Boughanem and J. Savoy,2008] "La pertinence est la correspondance entre un document et une requête, ou encore une mesure d'informativité du document à la requête".

Il existe plusieurs "'*pertinences*"' possibles entre un document et un besoin, nous en citons les quatre (3) les plus importantes :

- Pertinence algorithmique :Cette pertinence (appelée aussi pertinence système) est traduite par une mesure algorithmique basée sur le calcul de la pertinence de l'information par rapport `à la requête en utilisant des caractéristiques des requêtes d'une part et des documents d'autre part. C'est le seul type de pertinence qui est indépendant du contexte.

- Pertinence thématique: Cette pertinence est traduite par la *typicalité* et est définie par le degré de couverture du document du thème évoqué par le sujet de la requête. C'est le seul type de pertinence appliqué dans les campagnes d'évaluation tels que TREC.

- Pertinence cognitive :C'est la pertinence liée au thème de la requête, selon la perception ou les connaissances de l'utilisateur sur ce même thème ; cette pertinence est caractérisée par une dynamique qui permet d'améliorer la connaissance de l'utilisateur via l'information renvoyée au cours de sa recherche.

2.2 Système de recherche d'information(SRI) :

Le but principale d'un système de recherche d'information (SRI) est de retourner les documents pertinents "*les plus proches*" du besoin en information de l'utilisateur exprimé par une requête.

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 [Tambellini, 07].

Les systèmes de RI adoptent une démarche simple, couramment appelée : Processus en U de recherche d'information. Illustré dans la figure 1.1

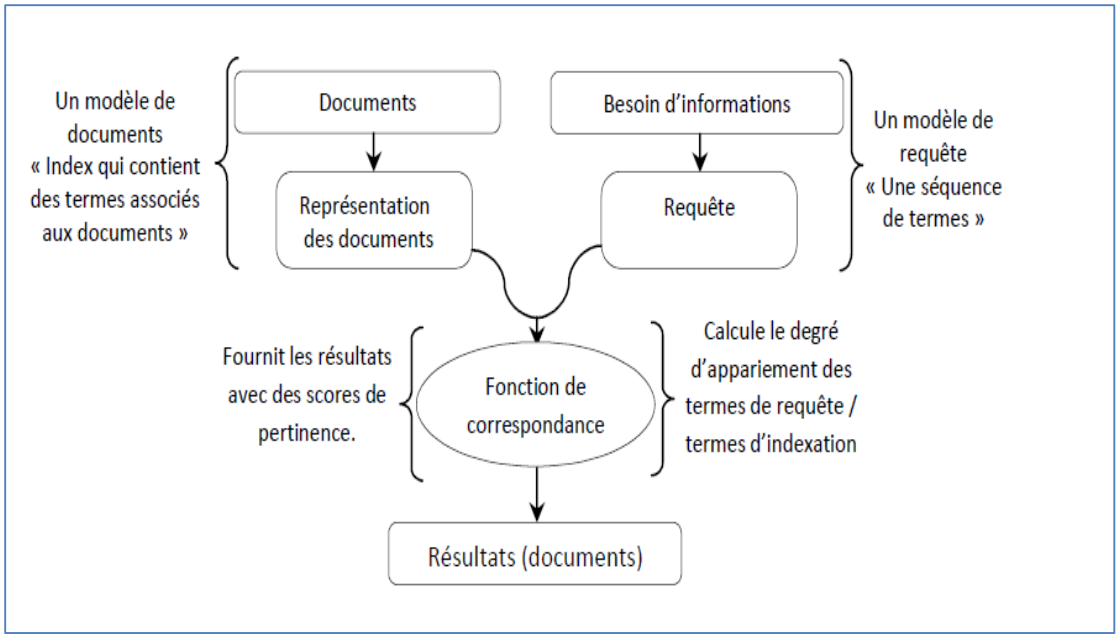


Figure 1.1 :Système de Recherche d'Information

2.2.1 Processus général de la RI :

Un système de recherche d'information transfert un corpus de documents en un corpus indexé à l'aide d'une fonction d'indexation. Requêtes traduites à partir de besoins utilisateur sont résolu par ce corpus. Le principe d'un tel système est basé sur la définition d'un modèle de recherche d'information qui effectue ces deux transpositions et qui fait correspondre les documents aux requêtes. La transposition d'un document en un document indexé repose sur un modèle de document. De même, la transformation du besoin utilisateur en requête repose sur un modèle de requête. Enfin, la correspondance entre une requête et des documents s'établit par une relation de pertinence [Maisonnasse, 08]. La figure 1.2, présente les différentes étapes d'un processus de recherche d'information.

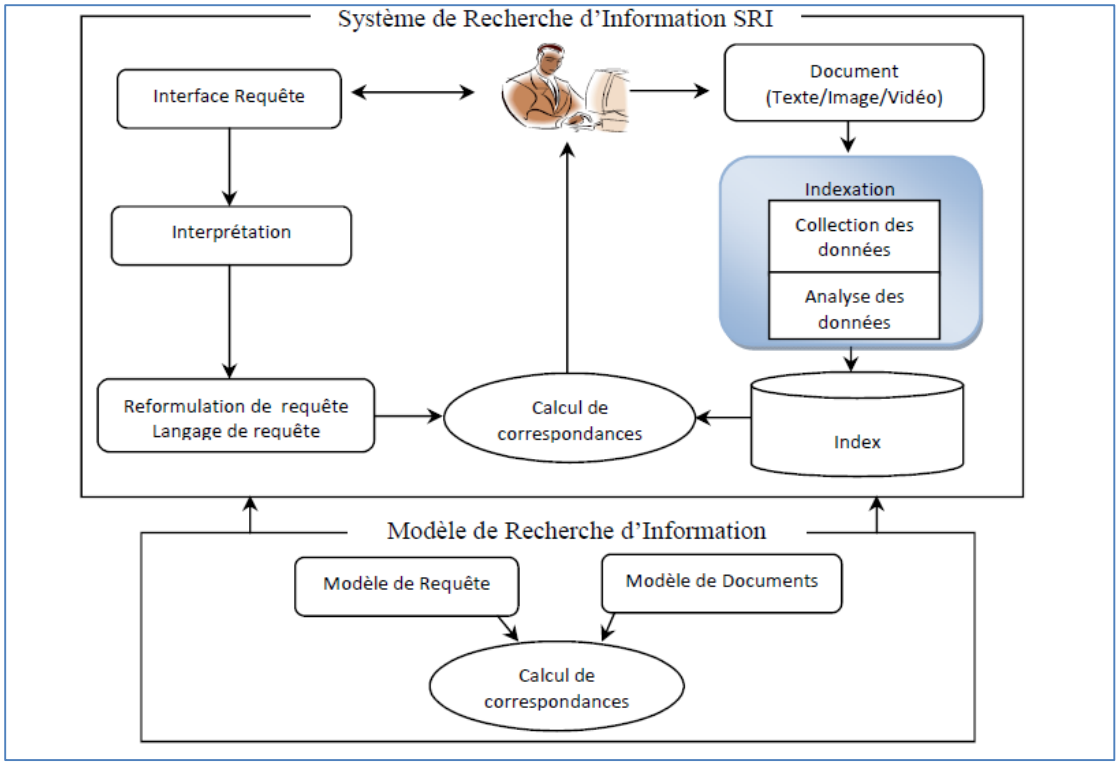


Figure1.2 :Processus de recherche d'information

2.2.1.1 La phase d'indexation :

L'indexation est une étape très importante dans le processus de RI. Cette dernière consiste à extraire des documents les mots les plus discriminants encore appelés index. Cette première tâche est généralement effectuée en marge du processus de recherche car, la construction des index peut être assez longue en fonction du nombre de documents de la collection ainsi que de la taille des documents. Les index ont un caractère réducteur car tous les termes d'un document ne sont pas importants à prendre en compte pour la recherche. L'indexation peut se faire de 3 manières différentes : manuellement (faite par un humain), de manière semi-automatique (par exemple créée par un humain assisté d'un programme proposant des termes), ou de manière automatique (crée par un programme informatique) [Kompaoré, 08].

- *Manuelle* : Chaque document est analysé par un spécialiste du domaine ou un documentaliste. Cette indexation présente un certain nombre d'inconvénients liés notamment à l'effort et le prix qu'elle exige (en temps et en nombres personnes). De plus, cette indexation est subjective, qui est liée au facteur humain, différents spécialistes peuvent indexer un document avec des termes différents.[hammech,2013]
- *Semi-automatique (mixte) :* C'est une combinaison des deux méthodes automatique et manuelle : un premier processus automatique permet d'extraire les termes du document. Cependant, le choix final reste au spécialiste du domaine ou au documentaliste pour établir les relations entre les mots clés et choisir les termes significatifs.[BOURAMOUL ,2011]
- *Automatique* : Chaque document est analysé à l'aide d'un processus entièrement automatisé. Dans ce cas, l'indexation est entièrement automatisée. Elle est réalisée par un programme informatique et elle passe par un ensemble d'étapes pour créer d'une façon automatique l'index. Ces étapes sont : l'analyse lexicale, l'élimination des mots vides, la normalisation (lemmatisation ou radicalisation), la pondération des mots et enfin la création de l'index et éventuellement sa compression. [hammech,2013]

Les techniques basiques utilisés dans l'indexation automatique mentionner précédemment sont détaillées comme suite :

- **L'analyse lexicale :**

Elle transforme un texte de document en une liste de termes. Un terme est un groupe de caractères constituant un mot significatif. L'analyse lexicale permet de reconnaître les espaces de séparation des mots, les chiffres, les ponctuations, etc.[hammech,2013]

- **Elimination des mots vides :**

Cette étape consiste à supprimer dans un premier temps des mots de fortes fréquences, généralement en référence à une liste de mots appelée anti-lexiques ou plus fréquemment *stoplist*. [Daoud,2009]

L'élimination des mots vides réduit la taille de l'index, ce qui améliore le temps de réponse du système.

- **La normalisation :**

La normalisation consiste à représenter les différentes variantes d'un terme par un format unique appelé lemme ou racine. Ce qui a pour effet de réduire la taille de l'index. Plusieurs techniques de normalisation sont utilisées : par exemple l'élimination des affixes [hammech,2013]

- **La pondération des termes :**

Cette étape est généralement basée sur des formules de pondération qui affecte à chaque terme un degré d'importance (une valeur de discrimination) dans le document où il apparait. Il existe un grand nombre de formules de pondération dont la plus connue est basée sur deux facteurs : fréquence de terme (TF) et fréquence inverse de document (IDF), définies dans ce qui suit :

TF (termfrequency) :

Cette mesure est proportionnelle à la fréquence du terme dans le document. L'idée sous-jacente est que plus un terme est fréquent dans un document, plus il est important dans la description de ce document.

Le *TF* est souvent exprimé selon l'une des déclinaisons suivantes :

1. *TF :* utilisation brute,

2. $1 + \log(f_{ti,dj})$

$f_{t,d}$: le nombre d'occurrence du terme ti dans le document dj

IDF (Inverse Document Frequency) :

Mesure l'importance d'un terme dans toute la collection. L'idée sous-jacente est que les termes qui apparaissent dans peu de documents de la collection sont plus

représentatifs du contenu de ces documents que ceux qui apparaissent dans tous les documents de la collection.

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

1. $IDF = \log \frac{N}{df}$

2. $IDF = \log \frac{N-df}{df}$

df : la proportion de documents contenant le terme ti

N : le nombre total de documents dans la collection.

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

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

2.2.1.2 La phase de recherche ou d'appariement document requête :

Le phase d'appariement document-requête sert à mesurer la pertinence d'un document vis-à-vis d'une requête. De manière globale, chaque fois l'utilisateur exécute une requête, le SRI calcule un score de pertinence (similarité vectorielle, similarité booléen, probabiliste, etc.). Ce score de pertinence est calculé selon une fonction ou une mesure de similarité, notée *RSV*(*Q*;*D*) (*Retrieval Status Value*) où *Q* est une requête et *D* un document de la collection. La phase d'appariement est fortement liée à la phase d'indexation et de pondération des termes. Il existe deux méthodes d'appariement :

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

2.2.2 Les modèles de recherche d'information :

Un modèle de RI fournit une interprétation théorique de la notion de pertinence. Il existe plusieurs modèles différents de RI qu'ont été utilisé dans cet environnement, ils se basent sur des domaines théoriques différents, théorie des ensembles, algèbre, probabilités, etc. généralement, on distingue trois principales classes de modèles : modèles booléens, modèles vectoriels et modèles probabilistes

2.2.2.1 Le modèle booléen :

- **Le modèle de base :**

Dans le modèle booléen en représentant la requête sous forme d'une équation logique. Le processus de recherche mis en œuvre par le système consiste à effectuer des opérations sur ensembles de documents définis par l'occurrence ou absence de termes d'indexation qui sont reliés par des connecteurs logiques *ET*, *OU* et *NON*.

$$RSV(D, q_i) = 1 \ si \ \ q_i \epsilon \ D \ ; 0 \ \ sinon.$$

$$RSV(D, q_i \wedge q_i) = 1 \ si \ \ RSV(D, q_i) = 1 \ \ et \ \ RSV(D, q_j) = 1; 0 \ sinon.$$

$$RSV(D, q_i \vee q_i) = 1 \ si \ \ RSV(D, q_i) = 1 \ \ ou \ \ RSV(D, q_j) = 1; 0 \ sinon.$$

$$RSV(D, \neg q_i) = 1 \ si \ \ RSV(D, q_i) = 0 \ ; 0 \ sinon.$$

Figure 1.3 :les règles de correspondances du model booléen

- **Le modèle booléen étendu :**

Considérons le cas d'une requête ne comportant que deux termes T1 et T2 et examinons le cas des requêtes conjonctives (T1 ET T2) et disjonctives (T1 OU T2), le but étant d'ordonner les documents en réponse à cette requête .

la similarité de la requête Q à un document est donnée par:

$$sim(q_{or}, d) = \sqrt{\frac{w_1^2 + w_2^2}{2}}$$

$$sim(q_{and}, d) = 1 - \sqrt{\frac{(1 - w_1)^2 + (1 - w_2)^2}{2}}$$

Les requêtes conjonctives et disjonctives peuvent être généralisées aux cas où la requête comporte plus de deux termes (termes). On utilise pour cela les p-normes qui dépendent d'un paramètre pouvant varier dans l'ensemble des entiers naturels. La généralisation de la similarité conjonctive (OU) s'exprime ainsi :

$$sim(q_{or}, D_j) = \sqrt[p]{\frac{w_1^p + w_2^p + \ldots + w_m^p}{m}}$$

Et la généralisation d'une requête disjonctive (ET) par :

$$sim(q_{and}, D_j) = 1 - \sqrt[p]{\frac{(1 - w_1)^p + (1 - w_2)^p + \ldots + (1 - w_m)^p}{m}}$$

- **Le modèle des ensembles flous :**

Cette extension du modèle booléen tient compte de la pondération des termes dans les documents. Le document est donc représenté par un ensemble de termes pondérés. La requête reste toujours une expression booléenne. L'évaluation de l'appariement d'un terme à une requête repose sur une des évaluations classiques proposées dans le cadre des ensembles flous :

RSV(Dj,q1∧q2) = min(RSV(Dj,q1), RSV(Dj ,q2))

RSV(Dj,q1∨q2) = max(RSV(Dj,q1), RSV(Dj ,q2))

RSV(Dj, ¬qi) = 1-(RSV(Dj, qi)

2.2.2.2 Le modèle vectoriel :

- Le modèle de base :

Ce modèle préconisé la représentation des requêtes utilisateurs et documents sous forme de vecteurs. Chaque document est représenté par un vecteur : dj = (w1j ,w2j , . . ,wMj). De même chaque requête est représentée par un vecteur : qi = (w1i,w2i, . . ,wMi). Avec : w correspond au poids d'un terme dans le document dj ou dans la requête qi.

Le mécanisme de recherche consiste donc à retrouver les vecteurs documents qui s'approchent le plus du vecteur requête.

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

Figure 1.4 :les mesures de similarités de modèle vectoriel de base

- Le modèle LSI :

L'objectif fondamental du modèle LSI [kechid,2009] est d'aboutir à une représentation conceptuelle des documents où les effets dus à la variation d'usage des

termes dans la collection sont clairement atténués. Ainsi, des documents qui obtiennent des termes co-occurrents ont des représentations proches dans l'espace défini par le modèle. La base mathématique du modèle LSI est la décomposition par valeur singulière SVD de la matrice Terme-Document. La SVD identifie un ensemble utile de colonnes de base de cette matrice. Ces vecteurs de base couvrent le même espace de vecteurs associés à la représentation des documents, car ils sont obtenus par rotation (multiplication par une matrice orthogonale) des vecteurs d'origine. En pose :

X : Matrice Terme-Document

T 0 : Matrice avec colonnes orthonormées qui couvre l'espace des colonnes de X

D 0 : Matrice avec colonnes orthonormées qui couvre l'espace des lignes de X

S 0 : Matrice diagonale formée des valeurs singulières qui résultent de la normalisation de T 0 et D 0

On a :

$X = T_0 S_0 D_0^T$

La propriété de la SVD pour le modèle LSI est qu'en raison du tri des valeurs singulières dans l'ordre décroissant, la meilleure approximation de *X* peut être comme suit :

$X = \sum_{i=1}^{k} T_{0xli} S_{0xli} D_{0xi}^T$

2.2.2.3 Le modèle probabiliste :

- Le modèle de base :

Le modèle probabiliste est fondé sur la théorie des probabilités**.** Il trie les documents selon leur probabilité de pertinence vis-à-vis d'une requête.

Le processus de recherche se traduit par calcul, de proche en proche, du degré ou probabilité de pertinence d'un document relativement à une requête. Pour ce faire, le processus de décision complète le procédé d'indexation probabiliste en utilisant deux probabilités conditionnelles : [zargayouna,2005]

- P(wij/Pert) : probabilité que le terme ti occurre dans le document Dj sachant que ce dernier est pertinent pour la requête.
- P(wij/NonPert) : que le terme ti occurre dans le document Dj sachant que ce dernier n'est pas pertinent pour la requête.

Le calcul d'occurrences des termes d'indexation dans les documents est basé sur l'application d'une loi de distribution sur un échantillon représentatif de documents d'apprentissage. En posant les hypothèses suivantes :

- La distribution des termes dans les documents pertinents est la même que leur distribution par rapport à la totalité des documents,
- Les variables « document pertinent » et « document non pertinent » sont indépendantes. La fonction de recherche est obtenue en calculant la probabilité de pertinence d'un document D, notée P(Pert/D) :

$$RSV\left(\frac{Pert}{D}\right) = \sum_{i=1}^{j} \frac{P\left(w_{ij}\middle|Pert\right)}{P\left(w_{ij}\middle|NonPert\right)}$$

- Le modèle de réseau inférentiel bayésien :

Un réseau bayésien est un graphe direct acyclique où les nœuds représentent des variables aléatoires et les arcs des relations causales entre les nœuds. Ces derniers sont pondérés par des valeurs de probabilités conditionnelles. Dans l'espace défini par les termes d'indexation, on définit :

- *T* : variables aléatoires binaires *t 1 ,..., t T* associés aux termes d'indexation
- *D j* : Variable aléatoire associée à un document
- *Q k* : Variable aléatoire associée à une requête

On calcule alors la mesure de pertinence de *Q k* relativement à *D j* en traitant les probabilités conditionnelles de Bayes selon la formule :

RSV(Q_k ,D_j)=1-P($Q_k \wedge D_j$)

Où :

$$P(Q_k \wedge D_j)=\sum_{i=1}^{T} P\langle Q_k|t_i\rangle * (\prod_{ti\in Dj} P\langle t_i|D_j\rangle * \prod_{ti\in Dj} P\langle T_i|D_j\rangle) * P(D_j)$$

Avec :

P (Q k / t i): Probabilité que le terme t i appartienne à un document pertinent de Q k

P (t i / D j): Probabilité que le terme t i appartienne au document D j sachant qu'il est pertinent

P (t i / D j) = 1- P (t i / D j)

P (D j): Probabilité d'observation D j

2.2.3 Evaluation des performances des systèmes de RI :

L'évaluation des systèmes de recherche d'information est donc un thème de recherche important en sciences de l'information. Elle peut porter sur plusieurs critères : le temps de réponse, la pertinence des résultats, la qualité de la présentation, etc. Le critère le plus important est sans doute celui qui mesure la capacité du système à satisfaire le besoin de l'utilisateur en information, cette satisfaction se traduit par une forte adéquation entre la requête émise et les documents retournés. Dans ce contexte des campagnes d'évaluation ont été mises en place depuis les années soixante pour juger l'efficacité de ces systèmes et ainsi faire évoluer leurs performances sur le plan technique mais également par rapport aux attentes des utilisateurs. [BOURAMOUL,2011]

2.2.3.1 Collection des tests :

Afin de mesurer le degré de pertinence des résultats d'un système de recherche d'information à une requête, il faut de connaître l'ensemble des documents pertinents qui réponds à une requête donnée par l'utilisateur. C'est à cette fin que des collections de tests ont été élaborées. Une collection de tests comprend : des corpus de documents, une liste de requêtes prédéfinies et des jugements de pertinence.

Corpus de documents :C ’est une collection de documents à indexer, il s’agit de collection des informations accessibles et exploitables sur lesquelles le système sera évalué. Un groupe d’experts dans un domaine participent généralement à la constitution d’un corpus documentaire homogène et cohérent couvrant le domaine. Dans le cas global et pour un souci d'optimalité, la base documentaire constitue des représentations simplifiées mais suffisantes pour ces documents. Des études sont fait sur ces représentations de telle sorte que l’ajout, la suppression d'un document ou la recherche de la base se font dans les meilleures conditions de coût.

Requête :C’est une représentation d’un besoin d’information. Il est souhaitable que le protocole d’évaluation prend en compte une requête directement dans la forme où elle a été soumise au système. Pour exploiter au mieux les caractéristiques de la collection de documents et avoir une évaluation assez objective, il est important de créer un ensemble de quelques dizaines de requêtes (TREC utilise de 25 à 50 topics comme une norme du nombre de requêtes de test), et qui soient adéquates par leur longueur, les thèmes abordés, leur forme, etc.[Ourdia,2011]

Jugement de pertinence : Ils sont manuellement créés et constituent la liste des documents pertinents pour chaque requête émis par l’utilisateur. Ils peuvent être portés par l’utilisateur lui-même, ou par un expert extérieur qui se concentrera plus facilement sur la « pertinence ». Ils peuvent prendre des formes booléennes variées (le document est pertinent ou il ne l’est pas), généralement par l’usage de pondérations.

2.2.3.2 Mesures d’évaluation de SRI :

L’objectif d’un système de recherche d’information est d’éliminer tous les documents non pertinents et de retourner à l’utilisateur seulement les documents pertinents. Cet dernière est évalué à l’aide de différentes mesures d’évaluation**.**

- **Rappel et précision :**Le rappel et la précision sont deux mesures de base pour évaluer les performances des systèmes de recherche d'information. Le principe de ces deux mesures est basé sur la partition de l'ensemble des documents restitués par le SRI en deux catégories : documents pertinents et documents non pertinents.

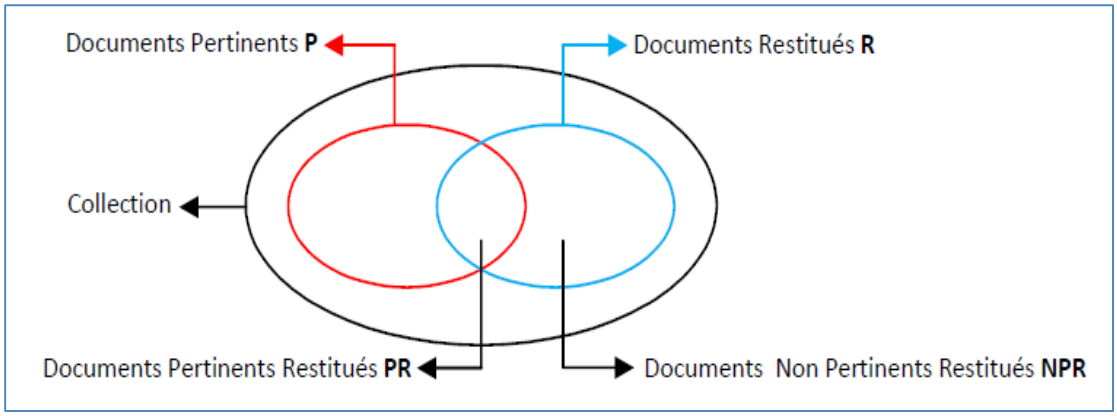


Figure 1.8:Partition de la collection pour une requête

➢ **La précision :**

Est le rapport du nombre de documents pertinents retournés par le système (SP) sur le nombre total de documents retournés (R)

Le principe est le suivant : quand un utilisateur interroge une base de données, il souhaite que les documents proposés en réponse à son interrogation correspondent à son attente. Tous les documents retournés superflus ou non pertinents constituent du bruit. La précision s'oppose à ce bruit documentaire. Si elle est élevée, cela signifie que peu de documents inutiles sont proposés par le système et que ce dernier peut être considéré comme « précis ». On calcule la précision avec la formule suivante :

$$\text{Précision} = \frac{SP}{R}$$

➢ **Le rappel :**

Est le rapport du nombre de documents pertinents retourné (SP) sur le nombre total de documents pertinents (P)

Cela signifie que lorsque l'utilisateur interroge la base, il souhaite voir apparaître tous les documents qui pourraient répondre à son besoin d'information. Si cette adéquation entre le questionnement de l'utilisateur et le nombre de documents présentés est importante alors le taux de rappel est élevé. À l'inverse, si le système possède de nombreux documents intéressants mais que ceux-ci n'apparaissent pas dans la liste des réponses, on parle de silence. Le silence s'oppose au rappel. Le rappel est donc calculé comme suit :

Le rappel = $\frac{SP}{P}$

- **Silence :**mesure l'ensemble des documents pertinents non sélectionnés (oubliés par le système) : **Silence= 1 – Rappel**
- **Bruit :**mesure l'ensemble des documents non pertinents sélectionnés (ajoutés par erreur par le système) : **Bruit= 1 – Précision**.

➢ **Courbe rappel-précision :**

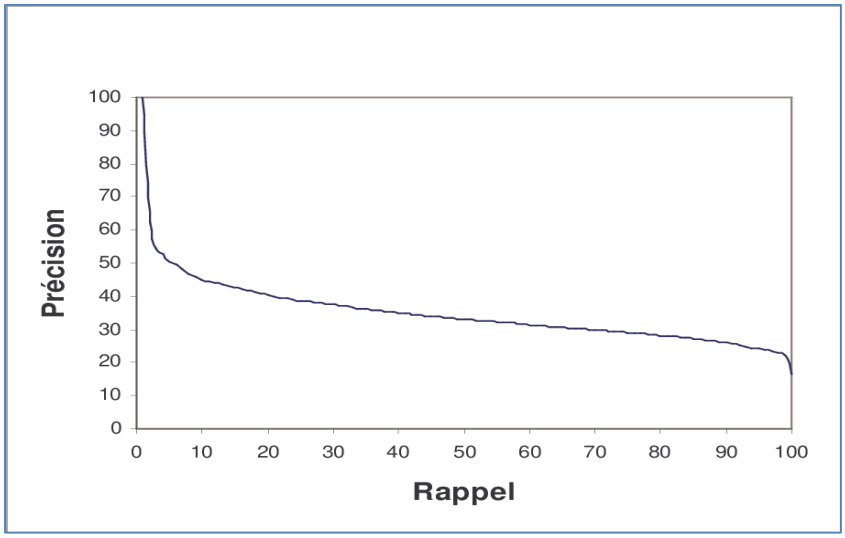


Figure 1.9:graphe d'une courbe rappel –précision

➢ **Quelques mesures sure le rappel-précision**

- **F-Mesure :** C'est une mesure qui combine la précision et le rappel, nommée F-mesure ou F-score définie par :

$$F_\beta = \frac{(1+\beta^2)*(precision * rappel)}{(\beta^2 * precision + rappel)}$$

Où β traduit l'importance relative du rappel et de la précision. La valeur maximale de la F-mesure est 1. Elle est obtenue quand tous les documents classés sont pertinents.

- **X-Précision :**C'est la précision à différents niveaux de coupe. Cette précision mesure la proportion des documents pertinents retrouvés parmi les X premiers documents retournés par le système.

- **R-Précision :**Cette précision mesure la proportion des documents pertinents retrouvés après que R documents ont été retrouvés, ou R est le nombre de documents pertinents pour la requête considérée.

- **MAP**(Mean Average Precision-MAP): Précision moyenne (MeanAveragePrecision-MAP) : c'est la moyenne des précisions moyennes (Averageprecision-AP) obtenues sur l'ensemble des requêtes `a chaque fois qu'un document pertinent est retrouvé [ourdia ;2011] :

$$\text{MAP} = \frac{\sum_{q \in Q} AP_q}{|Q|}$$

APq : est la précision moyenne d'une requête q
Q: est l'ensemble des requêtes
|Q| : est le nombre de requête

3. De la RI classique à la RI adaptative :

La limite principale de la RI classique est qu'elle est basée sur une approche générale qui évalue les requêtes des utilisateurs et délivrent des résultats sans tenir compte des critères spécifiques de l'utilisateur qui à exprimer la requête. Par exemple La requête " *jaguar*" est une requête ambigüe du fait qu'elle est liée à différentes thème de recherche, telles qu'une marque automobile anglaise jaguar, ou un l'animal ayant le nom *jaguar*. Dans ce type de requête le système doit retourner des résultats interdépendants aux centres d'intérêts de l'utilisateur lors de la recherche. [Daoud,2009]

L’objectif de la RI adaptative et d’adapter au besoin spécifique de l’utilisateur en termes des documents pertinents associé.

3.1 La RI adaptative :

Le but fondamental de la RI adaptative consiste à utiliser des informations supplémentaires, autres que la requête, extraites des interactions de l’utilisateur avec le système de recherche d’information, en utilisant des techniques développées pour désambiguïser le sens des mots de la requête utilisateur dans le but d’améliorer la pertinence de la recherche.

3.1.1 Reformulation de requêtes :

Pour mieux représenter les besoins en information de l’utilisateur, La reformulation de la requête consiste à construire une nouvelle requête par ajout de nouveaux termes à la requête initiale, suppression ou re-pondérer les termes dans le but de cibler la recherche vers les documents pertinents. La génération de la nouvelle requête s’effectue soit par reformulation automatique ou par reformulation interactive.

3.1.1.1 Reformulation automatique :

La reformulation automatique de requête ou *expansion de requête* est l'une des premières techniques ayant produit des améliorations notables dans ce domaine. L'idée de base est d'ajouter à la requête initiale des termes issus de ressources linguistiques existantes ou bien de ressources construites à partir des collections. [Zemirli,2011]

3.1.1.2 Reformulation manuelle :

On la nomme communément ré-injection de la pertinence ou relevance feedback. On présente à l’utilisateur une liste de documents retournées par le système comme réponse à la requête initiale. Après les Avoir vérifié, l’utilisateur indique ceux qu’il considère pertinents. L’idée Principale de la réinjection de pertinence est de sélectionner les termes Importants appartenant aux documents jugés pertinents par l’utilisateur, et de renforcer l’importance de ces termes dans la nouvelle formulation de la requête. [Ourdia,2011]

Pour utiliser cette technique on doit passer sur trois étapes principales :

- **L'échantillonnage :** cette étape permet de construire un échantillon de documents à partir des éléments jugés par l'utilisateur. Cet échantillon est caractérisé par le nombre d'éléments jugés et 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, la plus reconnue est celle de Rocchio [Rocchio, 1971] adaptée au modèle vectoriel.

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

3.1.2 Désambiguïsation du sens des mots de la requête :

Pour mieux exprimer le besoin en information relatif l'intention de recherche d'utilisateur, ce dernier adopte cette technique. Elle lui permet de saisir le vrai sens évoqué par les termes de sa requête. L'adaptation du processus de recherche `a des critères spécifiques de l'utilisateur par désambigüisation du besoin en information derrière la requête est adoptée dans plusieurs approches dans le domaine de la RI adaptative. La plupart de ces techniques se basent sur l'exploitation des interfaces de clarification interactives à base d'ontologie.

3.1.3 Regroupement thématique des résultats de recherche :

A cause de la croissance du web et les difficultés rencontrées par les moteurs de recherche d'information classiques pour satisfaire les besoins en information des requêtes formuler par l'utilisateur, 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.

3.2 Limitations de la RI adaptative :

Les limites sont principalement liées à la représentation limitée du contexte de l'utilisateur, le mode d'interaction explicite avec le système et à la dépendance de la familiarité de l'utilisateur avec le thème de recherche et le nombre d'itérations de recherche sur la performance de la RI adaptative.

4 La recherche d'information sémantique :

4.1 Définition :

La recherche sémantique d'information est une branche de la RI qui vise à étudier les significations des termes dans les documents pour permettre une meilleure compréhension du contenu de document et du sens de l'information afin de faciliter l'accès à l'information pour un utilisateur ayant un besoin en information.

La recherche d'information sémantique (RIS) a pour but de dépasser les limites d'une recherche classique par mots-clés. De plus en plus de travaux s'intéressent à l'exploitation de ressources sémantiques (ontologies, terminologies, thésaurus, etc.) pour améliorer l'accès à l'information.

4.2 Indexation sémantique :

L'indexation sémantique en RI est née du problème de l'ambiguïté des mots de la langue naturelle utilisés classiquement pour représenter les documents et requêtes. L'indexation sémantique a pour objet de représenter les documents et requêtes par les sens des mots permettant ainsi de lever toute ambiguïté, ce qui a pour conséquence d'améliorer les résultats de la recherche.

Pour retrouver les sens corrects des mots dans un document (ou dans une requête donnée), l'indexation sémantique utilise des techniques de désambiguïsation des sens des mots. Ces techniques se basent sur l'utilisation de ressources, telles que les corpus d'apprentissage ou les ressources terminologiques.

Nous décrivons, dans ce qui suit, les ressources linguistiques externes les plus exploitées par la désambiguïsation des sens des mots, puis nous présentons les principes de base de la désambiguïsation des sens des mots, ainsi que les travaux les plus significatifs dans le domaine. Les approches d'indexation qui sont basées sur les techniques de désambiguïsation linguistique des mots seront aussi présentées.

4.3 Les ressources linguistiques externes :

4.3.1 Les ressources structurées :

4.3.1.1 les dictionnaires informatisés (*Machine Readable dictionaries -MRDs-*) :qui représentaient des sources de connaissances très populaires dans les années 80 pour les différentes disciplines du domaine du traitement automatique de la langue. Dans un dictionnaire informatisé, un mot de la langue possède un ou plusieurs sens qui sont définis par leurs glossaires (*gloss*). Le glossaire d'un sens décrit le sens du mot par une définition, des commentaires et/ou des exemples d'utilisation courante

4.3.1.2 les thésaurus :sont des vocabulaires contrôlés permettant de définir les termes et les relations sémantiques entre termes (*Hyperonymie/Hyponymie, Synonymie, Antonymie, …etc.*). Les termes dans un thésaurus sont organisés dans des catégories de domaines décrivant leurs domaines d'usage.

4.3.1.3 les ontologies :sont des spécifications conceptuelles d'un domaine. L'ontologie définit et organise l'ensemble des concepts du domaine et les relations sémantiques entre eux (taxonomie, synonymie, …).

4.3.1.4 Taxonomie : Une taxonomie est un vocabulaire contrôlé organisé sous une forme hiérarchique simple. Les liens hiérarchiques dans une taxonomie correspondent à des liens de spécialisation affinant le sens d'un terme. La taxonomie, à la différence du thesaurus ne permet pas de parcourir la hiérarchie de manière connexe ni de restreindre ou de spécialiser le champ des connaissances.

4.3.2 Les ressources non structurées :

4.3.2.1 les corpus d'apprentissage : Ce sont de longs textes utilisés dans les techniques d'apprentissage pour construire la connaissance nécessaire pour la désambiguïsation des sens des mots (WSD). Ces corpus peuvent être étiquetés manuellement avec les sens des mots.

4.3.2.2 les corpus de collocations : Ce sont des ensembles de collocations de mots qui ont une tendance de se produire ensemble régulièrement.

4.4 Les approches d'indexation sémantique basée sur la désambiguïsation :

L'indexation sémantique a pour but de représenter les documents et les requêtes par les sens des mots qu'ils contiennent plutôt que par les mots eux-mêmes. Les sens des mots sont identifiés par des techniques de désambiguïsation qui s'appuient soit sur des ressources linguistiques externes telles que les dictionnaires, les thésaurus, les ontologies, …etc, soit sur des corpus d'apprentissage. Par conséquent, deux types d'approches d'indexation sont à distinguer : les approches d'indexation basée sur les ressources linguistiques externes et les approches d'indexation basées sur les corpus d'apprentissage.

4.4.1 La méthode de Voorhees :

[Voorhees ,1993] **Voorhees** a créé un outil particulier à la désambiguïsation qui repose sur le terme *WordNet*. Ce terme est un réseau utilisé dans le domaine sémantique et structuré concernant le concept *synset*, Un *synset* consiste *à* rassembler des termes que ce soit simples ou composés ayant le même sens dans un contexte donné.

Pour supprimer l'ambiguïté d'un mot, les *synsets* de ce mot calculent la valeur entre le contexte de ce mot et un voisinage contenant les mots du *synsets* dans la hiérarchie de

WordNet.

Après l'application de cette approche sur l'ensemble de tests de désambiguités effectué par Voorhees, ils ont trouvé que l'outil de désambiguïsation était incapable de déterminer de manière exacte le sens attendu des mots.

4.4.2 La méthode de Krovetz & Croft :

[Krovetz et Croft,1992] Dans les années 1992, Krovetz et Croft faire une grande étude qui permet de connaitre la relation de correspondance du sens des mots dans les requêtes et les documents. Leur but était de trouver l'existence d'une relation entre la correspondance/non correspondance des sens.

4.5 Les approches de désambiguïsation des sens des mots (WSD) :

L'objectif principal de la WSD est de retrouver les sens corrects des mots dans leur contexte d'utilisation. De nombreuses approches de désambiguïsation sémantique des mots existent.

- Lesk [Lesk, 1986] a développé l'un des premiers systèmes de désambiguïsation basé sur un MRD. Dans ce système, les sens possibles d'un mot cible sont d'abord identifiés à partir d'un dictionnaire. Un score est ensuite associé à chacun de ces sens. Le score associé à un sens possible correspond au nombre de mots communs entre la définition (ou gloss) dans le dictionnaire, et les définitions des sens des autres mots co-occurrents dans son contexte (le contexte est ici défini comme une fenêtre de dix mots autour du mot cible). Le sens qui maximise ce score est retenu comme sens correct du mot cible.

- Wilks et al. [Wilks et al., 1990] ont repris le principe de Lesk, en proposant d'étendre le contexte local d'un mot ambigu et ses entrées-sens pour obtenir un maximum de recouvrement de mots entre eux. La technique d'expansion utilisée consiste à enrichir manuellement le contexte d'un mot et les définitions de ses sens avec des mots qui co-occurrent fréquemment avec les mots de leurs textes respectifs (contexte et sens). Le contexte et les différentes entrées-sens du mot cible élargis sont ensuite représentés par des vecteurs. Le sens dont le vecteur associé a le plus grand degré de corrélation avec le vecteur contexte est retenu comme sens correct du mot ambigu.

5 CONCLUSION :

Nous avons présenté au cours de ce chapitre les principales notions et concepts de la recherche d'information telle que la requête, le document et la pertinence. Nous avons aussi décrit les processus de base de la RI, à savoir l'indexation, la pondération des termes et l'appariement document-requête. Ensuite, nous avons étudié les différents modèles de la RI ainsi que l'évaluation de performance des systèmes de recherche d'information. Et ensuite l'évolution de la RI classique vers la RI adaptative où nous avons citée, les techniques développées en RI adaptative, notamment la reformulation des requêtes, la désambiguïsation du sens des mots des requêtes et le regroupement thématique des résultats. Compte tenu des limitations de la RI adaptative, les travaux se sont orientées vers la RI contextuelle dont le but est de mieux répondre aux besoins en informations de l'utilisateur. Et enfin on la recherche sémantique ou nous avons citée quelque approche d'indexation.

Chapitre 2 :

La recherche d’information personnalisée

Chapitre2 : recherche d'information personnalisée

1. INTRODUCTION :

La RI adaptative présente des limitations principalement liées à la représentation de l'utilisateur et de ses besoins. Cette limitation a conduit à l'émergence de la RI personnalisée. Ce domaine apporte principalement la prise en compte de l'utilisateur en tant que composante principale dans le processus de la recherche.

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

La personnalisation a pour objectif, d'une part, de faciliter l'expression du besoin utilisateur et de rechercher des informations sur un sujet en écartant l'information non pertinente et donc réduire l'espace de recherche et d'autre part, de rendre cette information sélectionnée intelligible à l'usager et exploitable.

Plusieurs approches de personnalisation ont été définies dans la littérature. Quel que soit l'approche de personnalisation, on a toujours besoin de collecter et sauvegarder des données décrivant les utilisateurs sous forme de ***profils***. Il convient de distinguer la notion de *profil* de la notion de *requête*. Un profil peut être défini comme une mise en équation du centre d'intérêts et des préférences de l'utilisateur, alors qu'une requête est l'expression d'un besoin circonstancié que l'utilisateur souhaite voir satisfait.

Dans ce chapitre on va définir les notions de base de la recherche d'information personnalisée puis on explique les services de personnalisation les plus couramment utilisés ensuite en décrit le système de recherche d'information personnalisé avec ces concepts détaillés, après on va discuter l'évaluation des systèmes d'accès personnalisé à

l'information, en fin on va citer quelque approche pour la représentation sémantique de profil dans la recherche personnalisé et en finissant avec une conclusion.

2. Notions de base :

2.1 Contexte de recherche :

Il a été largement admis que la troisième génération des systèmes de recherche de l'information doit prendre en considération le contexte de recherche des utilisateurs dans tout le processus d'accès à l'information. [Kechid,2009].

La notion de contexte n'a pas une définition complète et collectif dans la littérature car cette notion est assez difficile à définir. D'une manière globale, le contexte regroupe des facteurs de déférentes natures qui limitent la compréhension, le champ d'application ou les choix possibles. Les facteurs plus fréquemment citées concernent des données spatio-temporelles (lieu, Heure, jour.) Ou des connaissances spécifiques en relation avec le domaine étudié des données culturelles.

2.2 Pertinence contextuelle :

La notion de pertinence contextuelle dépendant absolument du centre d'intérêt de l'utilisateur, du lieu et du support que l'utilisateur a choisi pour accéder à l'information.[Zemirli,2009]

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

2.2.1 Pertinence cognitive : c'est la pertinence liée au thème de la requête, « pondérée » par la Perception ou les connaissances de l'utilisateur sur ce même thème.

2.2.2 Pertinence situationnelle : c'est la pertinence liée à la tâche de recherche. Ce type de pertinence traduit essentiellement l'utilité de l'information relativement au but de recherche de l'utilisateur.

2.3 Profil utilisateur :

Le terme de profil utilisateur a été largement abordée dans le domaine du *user modeling*. Depuis le début des années 70, Le concept de profil utilisateur a été introduit

pour l'accès à l'information pour décrire une structure représentative de l'utilisateur, en présentant ses centres d'intérêts.

On dit « profil utilisateur toute structure qui permet de modéliser et de stocker les données caractérisant l'utilisateur ». Ces données représentent les centres d'intérêts, les préférences et les besoins en informations de l'utilisateur ou un groupe d'utilisateurs. [Zemirli,2009]

Un profil utilisateur est défini comme « une source de connaissance qui contient des acquisitions sur tous les aspects de l'utilisateur qui peuvent être utiles pour le comportement du système »

Ces définitions mettent en évidence trois aspects du profil qui s'exploitent ainsi : [Bouramoul,2011]

2.3.1 Source de connaissance :

Le profil utilisateur peut contenir des informations très diverses selon la tâche à accomplir, en recherche d'information le contenu d'un profil utilisateur se résume : ses caractéristiques personnelles, ses intérêts et ses préférences, ses compétences, son but courant et enfin l'historique de ses interactions avec le système. Le contexte contient des informations complémentaires permettant une meilleure

2.3.2 Acquisition :

Le contenu du profil utilisateur est une connaissance à récupérer, selon le degré d'adaptation du système, les données du profil utilisateur peuvent être soient, renseignées par l'utilisateur lui-même, soient récupérées par sélection d'un profil préexistant créé par des experts du domaine, ou encore capturées par le système de recherche d'information au cours de l'utilisation.

2.3.3 Utile pour le comportement du système :

En recherche d'information l'apport du profil utilisateur est de permettre une personnalisation ou une adaptation des services pour améliorer les performances du système, ou encore pour filtrer les résultats retournés par un moteur de recherche.

3 Les services de personnalisation :

Dans cette section on va présenter les services de personnalisation les plus couramment utilisés qui sont : Le filtrage du résultat, le re-ordonnancement des éléments du résultat, la recommandation d'éléments, et l'extension des requêtes avec les intérêts de l'utilisateur pour permettre la prise en compte de la personnalisation [Bouramoul,2011].

3.1 Le filtrage des résultats :

Le filtrage est généralement interprété comme l'élimination des informations qui ne répond pas au besoin de l'utilisateur sur un flux entrant, plutôt que la recherche des informations spécifiques sur ce flux. [Bouhini,2014].

Dans la technique de filtrage, un profil utilisateur doit décrire les ce qu'attire l'utilisateur. Un tel profil est alors comparé aux documents entrants afin d'essayer de déterminer ceux qui pourraient être intéressants pour cet utilisateur particulier. A l'arrivée de nouveaux documents, le système utilise le profil utilisateur pour choisir les documents pouvant potentiellement l'intéresser. Par un processus de retour de pertinence (« Relevance feedback », l'utilisateur n'indique pas seulement les documents pertinents mais aussi les documents non-pertinents (figure 2.1). Le système utilise ces informations pour ajuster la description du profil utilisateur qui reflètera les nouvelles préférences. [bouhini,2014].

Le principal avantage du filtrage des résultats est sa simplicité parce qu'il ne nécessite aucune modification du fonctionnement des fournisseurs d'information. Tout le traitement est fait après l'exécution de la requête. Les inconvénients sont le volume de données échangées entre le serveur et le client et le risque d'élimination d'éléments pertinents.

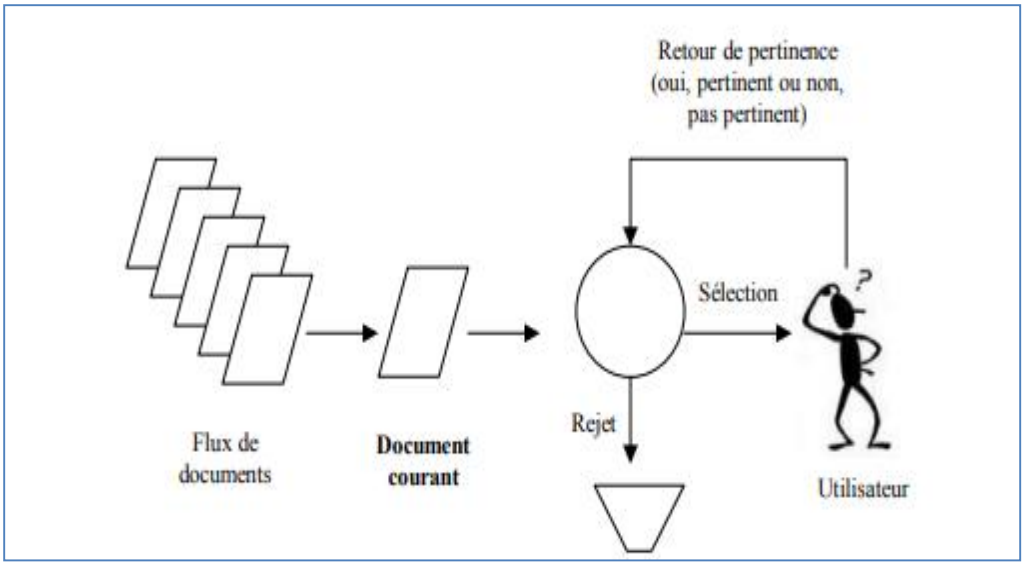


Figure 2.10:Représentation d'un système de filtrage

Un exemple de système qui filtre les résultats une fois la requête exécutée est CASPER (Case-BasedProfiling for ElectronicRecruitment). [Bouramoul,2011]

3.2 Le re-ordonnancement des résultats de la requête :

Le principe du re-ordonnancement est de modifier l'ordre de l'affichage des résultats à l'utilisateur. 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 sans pour autant négliger l'ordre qui a été attribué aux objets du résultat par le moteur de recherche. L'échange de l'ordre d'apparition des éléments du résultat est fait généralement en appliquant une fonction qui permet de calculer le nouveau rang de l'objet. Un exemple de fonction permettant d'effectuer le re-ordonnancement est donné par. [Bouramoul,2011]

Le re-ordonnancement des résultats de la requête a le grand avantage de ne pas exclure des éléments du résultat. C'est en quelque sorte une garantie que l'utilisateur va trouver ce qui l'intéresse et le seul objectif de cette approche est de faire gagner du temps aux utilisateurs. C'est une technique très peu utilisée dans les systèmes de personnalisation parce qu'elle nécessite le calcul du nouveau rang de chaque élément du résultat ce qui demande beaucoup de temps.

3.3 Reformulation de la requête :

Reformuler une requête a pour objectif de liée le langage de recherche naturel (utilisateur) et le langage d'indexation des systèmes de recherche d'information. Cela consiste à produire une nouvelle Requête en utilisant d'autres données. Pour la personnalisation, les données utilisées sont principalement relatives à l'utilisateur. Il s'agit alors d'un enrichissement de la requête. Dans ce cadre, les données intégrées sont les éléments des centres d'intérêt ainsi que des préférences. En 2003 Gauch et al. ajoutent à la requête les préférences implicites de l'utilisateur. En 2004 Li and Zaiane, déduisent les différentes catégories des profils stockés. Dans leurs travaux en bases de données. En 2005 Messai et al. enrichissent les requêtes en leur ajoutant de nouvelles propriétés à partir des ontologies de domaine disponibles. [Hadjouni Krir,2012]

3.4 La recommandation :

Le but d'un système de recommandations est d'aider les utilisateurs à faire leurs choix dans un domaine où ils disposent de peu d'informations pour trier et évaluer les alternatives possibles. Un système de recommandations peut être décomposé en trois entités de base (figure 2.2) : le groupe d'agents producteurs de données brutes, le module de calcul de recommandations et le groupe de consommateurs des recommandations. [Bouramoul,2011]

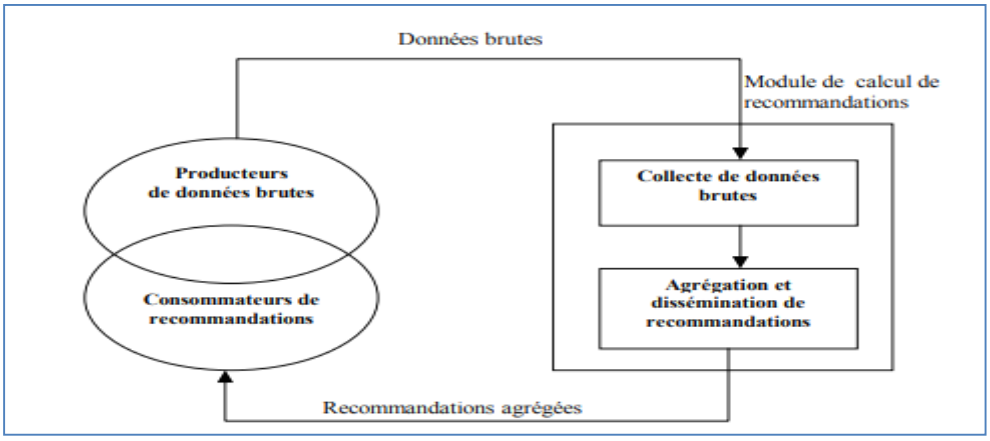


Figure 2.2 : Architecture générale d'un système de recommandation

Un autre type de recommandation consiste à formuler des propositions en fonction du profil de l'utilisateur et du contenu visualisé par ce dernier. C'est notamment cette

technique qui est utilisée par Amazon.com pour conseiller ses clients : si le visiteur s'intéresse à un livre, le site lui propose les ouvrages ayant également plu aux acheteurs de ce même livre. Pour Amazon, ce système vise à remplacer le travail que ferait un vendeur dans une librairie traditionnelle

4 System de recherche d'information personnalisée :

4.1 Définition :

Un système d'information personnalisé est un système qui intègre l'utilisateur, tout au long de la chaîne d'accès à l'information. Un tel système inclut alors :

- Des structures permettant de représenter l'utilisateur. Ces structures traduisent essentiellement les centres d'intérêts, les attirances de l'utilisateur
- Des techniques pour collecter les informations descriptives de l'utilisateur.
- Un processus d'accès à l'information intégrant les structures descriptives des utilisateurs.
- Un mécanisme d'évolution de ces structures.

4.2 Architecture générale (SRIP) :

La diversité des systèmes de personnalisation rend la définition d'une architecture fonctionnelle et formelle d'accès personnalisé à l'information difficilement généralisable. Néanmoins, L'architecture standard dégagée d'un système de Recherche d'information personnalisé (SRIP) peut être présentée par la figure suivante (figure 2.3).

Cette architecture est centrée autour de l'utilisateur. En ce sens l'ensemble des modules du système fait intervenir en partie les informations descriptives du profil utilisateur.

Sur la base de cette architecture, deux fonctions fondamentales sont dégagées : la gestion des profils et la sélection de l'information. [Lechani et Boughanem 2005]

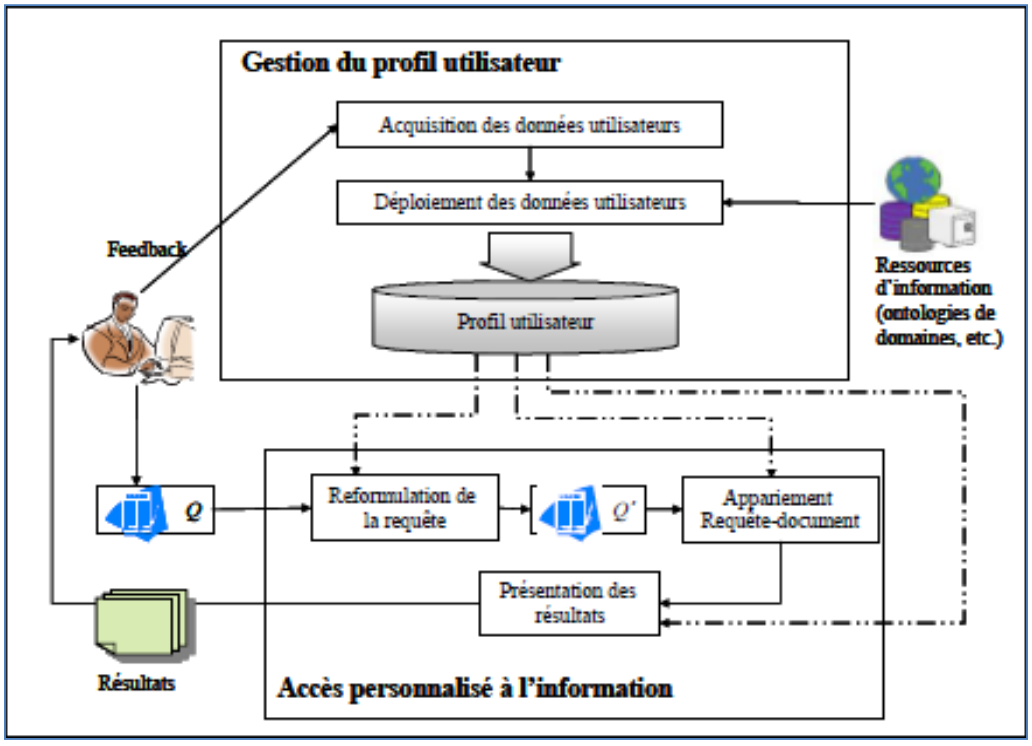


Figure 2.3 : Architecture générale d'un SRIP

4.2.1 Gestion des profils *:*

4.2.1.1 Représentation du profil utilisateur :

La modélisation de l'utilisateur est un processus à différentes étapes à savoir, la représentation du profil utilisateur qui peut être [karine abbas2008] :

- **Représentation vectorielle :**

Cette approche s'appuie généralement sur le modèle vectoriel. Le profil est formalisé sous forme de vecteurs des termes pondérés ou de classes de vecteurs. Le contenu est constitué d'un ou de plusieurs vecteurs visés dans un espace de termes obtenus implicite ou explicite à partir de plusieurs sources d'information. Les coordonnées des vecteurs correspondant aux poids des termes dans le profil. L'utilisation des avantages et des vecteurs permet de profiter de la diversité des domaines d'intérêts et gérer leur évolution dans le temps. Cette approche vous propose l'avantage de la simplicité de mise en

œuvre. Cependant les modèles proposés dans le passé, dans la dimension des temples de l'évolution des profils, ni à l'organisation des informations pour hiérarchiser les centres d'intérêts. Cette représentation manque de structuration et facilite ni l'interprétation du prix dans les différentes langues du pays caractérisant l'utilisateur. En effet, il n'y a pas de très variés et n'ont pas le même degré d'importance pour chaque utilisateur.

- **Représentation sémantique :**

La représentation sémantique rencontrée en relief les relations entre le contenu des informations reprenant le profil apport de solutions aux problèmes de dissection et de synonymie. Type CE de représentation utilisé généralement des ontologies ou de réseaux sémantiques probabilistes. Cette approche se base sur la construction hiérarchique de concepts à partir des questions d'information des figors*logs* . Chaque catégorie de la hiérarchie représente la connaissance d'un domaine d'intérêt de l'utilisateur.

- **Représentation booléenne et booléenne étendue :**

Cette approche est basée sur la présence ou l'absence des termes de la demande dans les représentations des documents (termes d'indexation associés). Le permet d'obtenir une recherche stricte réalisée à partir d'une comparaison exacte entre le besoin en informations, décrites dans la requête à l'Aide d'opérateurs logiques (ET, OU, NOT ou SAUF) et les termes représentant les documents d'une collection. Ce modèle a pour inconvénients de ne pas pouvoir ordonner les documents retournés. D'en plus, l'écriture de la requête peut être complexe et une mauvaise formulation peut engendrer des résultats attendus. Une extension du modèle booléen a été proposée pour essayer de palier le premier inconvénients (ordonnancement des résultats) en s'appuyant sur la théorie des ensembles flous. Le modèle étendu à prendre en plus des absences ou des présences des requêtes dans le document, les poids sémantiques de chacun de ces termes. L'inconvénient majeur de cette méthode de recherche est que l'évaluation peut proposer des valeurs différentes pour des expressions équivalentes.

- **Représentation multidimensionnelle :**

Cette représentation se base sur le principe que le modèle utilisateur doit contenir toutes les dimensions qui représentent l'utilisateur. L'idée de cette représentation est de diviser le profil utilisateur dans plusieurs catégories du système à l'environnement de l'utilisateur ou du niveau d'expertise. Plusieurs travaux ont abordé cet aspect sans le couvrir dans son ensemble. Chaque catégorie contient des informations spécifiques. Le profil dans cette approche est représenté par un ensemble de dimensions, distinguant principalement huit dimensions : les données personnelles, le domaine, le centre d'intérêts, la qualité des résultats, la personnalisation, la confidentialité, le feedback, et les informations générales. [Kechid,2009]

- **Représentation conceptuelle :**

La représentation conceptuelle est basée sur l'exploitation des ontologies de domaine ainsi que sur des hiérarchies. Les centres d'intérêt de l'utilisateur sont représentés sous forme de réseau de nœuds conceptuels reliés entre eux en suivant la topologie des liens définis dans les hiérarchies et les ontologies de domaines [Daoud, 2009].

Chacun des concepts est représenté par un vecteur de termes pondérés, la pondération traduit le degré d'intérêt de l'utilisateur. La représentation conceptuelle peut être considérée semblable à la représentation sémantique de par le fait que les centres d'intérêts de l'utilisateur sont représentés par un réseau de nœuds conceptuels. De même, cette représentation peut être assimilée à une approche ensembliste. En effet, ceci revient au fait que les domaines y sont généralement représentés en vecteurs de termes pondérés [Zemirli,2009].

4.2.1.2 Construction du modèle de l'utilisateur :

La construction du modèle de l'utilisateur repose sur deux phases. La première phase est la collecte d'informations sur l'utilisateur et la seconde est la construction proprement dite du modèle. Vient ensuite une phase d'évolution du modèle. Nous

présentons dans cette section les techniques d'acquisition des données utilisateur ainsi que les techniques de construction du modèle. Ensuite nous introduisons la notion de gestion de l'évolution du modèle de l'utilisateur.

Construire un modèle de l'utilisateur nécessite la connaissance de la structure à mettre en place afin de stocker les informations le concernant et de les exploiter d'une manière optimale. Cette dernière se base sur des approches et des techniques qui sont détaillées ci-dessous :

- **Acquisition des données utilisateurs**

La phase d'acquisition des données utilisateur consiste à collecter les informations qui serviront à l'instanciation du modèle de l'utilisateur. Nous distinguons deux modes d'acquisition : acquisition explicite et acquisition implicite. Dans l'approche explicite, les informations sont directement obtenues de l'utilisateur. Dans l'approche implicite, ce sont les données de comportement de l'utilisateur qui sont exploitées. Nous détaillons ces deux approches dans ce qui suit.

- **Acquisition explicit:**

Pour cette technique, l'utilisateur est directement interrogé à travers des formulaires à remplir, par exemple. Elle exploite des informations issues et fournies par l'utilisateur lui-même. Ces données peuvent être :

- Des jugements de pertinence pour les documents visités
- Des mots clés représentatifs des centres d'intérêts de l'utilisateur
- La sélection de liens ou de thèmes favoris

Ce type d'acquisition est utilisé dans plusieurs systèmes de RI personnalisée en ligne. Dans Google personalized search version 1.1, l'utilisateur peut saisir un ensemble de catégories représentatives de ses centres d'intérêt.

Ce mode d'acquisition est caractérisé par la simplicité de sa mise en place. Il présente le principal inconvénient d'être couteux en temps d'exécution. Un second inconvénient est

le décalage pouvant exister entre l'intention de l'utilisateur et ce qu'il désire réellement au moment de sa recherche.

- **Acquisition implicite :**

Dans cette technique de collecte de données, l'utilisateur n'est pas directement impliqué. En effet, les données utilisées sont issues de l'observation des comportements et des interactions de l'utilisateur avec le système lors de ses recherches.

Les comportements de navigation sont observés, enregistrés et ensuite traités. Parmi ces traces, nous citons :

- L'historique des visites ainsi que leur fréquence : ici, le temps passé sur une page visitée peut aussi être pris en compte. Il est à noter que ce temps est généralement comptabilisé au-delà d'un seuil minimum.
- Les cliques et les défilements de la souris sur les différentes pages parcourues. L'ordre d'impression, de sauvegarde ou de copie d'un document peuvent aussi intervenir.

Cette méthode implicite présente le principal avantage d'être transparente pour l'utilisateur, elle ne nécessite aucune effort d'attention particulier lors de sa recherche, ni de temps passé à émettre des jugements mais a néanmoins certains inconvénients : elle peut être moins précise que l'évaluation explicite par l'utilisateur mais cependant beaucoup plus dynamique dans le temps.

Le procédé d'acquisition implicite du profil d'utilisateur peut être classifié comme suit :

- **Un modèle peu profond** basé sur l'observation du comportement d'interaction relativement à court terme avec un système ; il ne tient pas compte des interactions del'utilisateur avec le système durant les sessions précédentes.

- **Un modèle profond** qui observe le comportement de l'utilisateur durant son interaction à long terme avec un système. Il tient compte de l'historique entier des interactions.

- **Techniques de construction :**

Les modèles des utilisateurs sont construits en utilisant les sources d'information utilisateur. A cet effet, un ensemble de techniques de construction basées sur l'apprentissage ou sur l'extraction d'information est utilisée.

Les techniques les plus courantes utilisées dans la construction de modèles à base de mots-clés ou de concepts sont détaillés comme suite :

- **Extraction d'ensemble de termes :**

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

- **Extraction de réseaux de termes :**

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

- **Extraction de concepts :**

Cette technique de construction concerne principalement les modèles d'utilisateurs représentés par une hiérarchie de concepts pondérés. Le principe de base de cette technique est l'utilisation d'une ta0xonomie de concepts de référence comme profil de base. Cette dernière peut être aussi bien l'ODP du projet Open Directory Project, un annuaire de concepts hiérarchique open source ou encore la hiérarchie de concepts Yahoo. L'approche de construction présente de manière générale les étapes deux suivantes : (1) identification des concepts et niveaux de l'ontologie à exploiter et (2) extraction des centres d'intérêts de l'utilisateur par analogie aux concepts de l'ontologie. Des systèmes tels que Persona (approche de coloration d'arbre), ARCH (approche hybride combinant vecteurs de termes et hiérarchie de concepts) et le système du projet OBIWAN (association des documents collectés avec les nœuds de l'ontologie).

4.2.1.3 Evolution du modèle de l'utilisateur :

Une des caractéristiques des systèmes de personnalisation est la gestion de l'évolution du modèle de l'utilisateur. Cette gestion consiste en premier lieu à capturer les transformations des centres d'intérêt de l'utilisateur et en second lieu à propager ces changements au niveau de la représentation du modèle de l'utilisateur. L'évolution du modèle peut s'effectuer lors de sa modélisation par deux modèles, l'un à court terme et l'autre à long terme ou encore en appliquant des algorithmes d'évolution génétique. L'adaptation du modèle de l'utilisateur implique d'avoir des changements dans les centres d'intérêts de l'utilisateur ce qui peut comprendre aussi la suppression ou l'ajout de domaines. En effet, faire évoluer le contenu du modèle de l'utilisateur consiste à adapter aussi bien la structure que le contenu aux évolutions des centres d'intérêts de l'utilisateur. Peu de travaux ont abordé cet aspect du modèle de l'utilisateur [Zemirli, 2008].

4.2.2 Sélection de l'information *:*

Cette phase consiste à inclure le profil ou le modèle utilisateur construit dans le processus de recherche d'information. En ce sens, les informations contenues dans le profil courant sont exploités pour identifier éventuellement le profil parmi ceux qui sont en cours de

construction, réécrire puis exécuter la requête, enfin présenter les résultats de la recherche. [Lechani et boughanem 2005].

4.2.2.1 Identification du profil :

C'est une opération qui consiste à apparier la structure instanciée du profil avec ceux décrits ou construits précédemment dans le système. L'appariement est généré sur le calcul d'un score de prévisibles de prédiction de profil ou similarité avec une classe de profils.

4.2.2.2 Exécution des requêtes :

L'exécution d'une requête traduite la succession éventuelle des opérations de sélection d'information, reformulation et calcul d'un score de pertinence. La sélection personnalisée d'information est une pratique courante dans les méta-moteurs de recherche. Leur principe est identifié, à travers le profil utilisateur, le type de requête (besoin général, actualité, scientifique ...) puis l'adresser à un moteur de recherche appropriée afin d'augmenter la précision des résultats. La reformulation de requête qui est une des premières techniques qui s'apparentent à la personnalisation, consiste à augmenter la requête avec des informations issues du profil avant de lancer le processus d'appariement. La personnalisation peut également porter sur la définition de la fonction de calcul de la pertinence.

4.2.2.3 Présentation des résultats :

La présentation des résultats est la phase finale du processus d'accès à l'information. Cette phase peut également considérer le profil de l'utilisateur en réordonnant les résultats fournis par le processus de sélection. En ce sens que l'ordre final des documents à présenter à l'utilisateur est une combinaison de l'ordre produit par le processus de sélection et celui donné par le contexte de l'utilisateur via un calcul de similarité ou jugements explicites de la pertinence. [Kechid,2009]

4.3 Les prototypes des systèmes d'accès personnalisés :

Les systèmes de recherche d'information personnalisés, sont devisés en trois Catégories principales qui sont : les systèmes de recommandation, les systèmes d'accès contextuel et les méta-moteurs de recherche personnalisée, plusieurs systèmes d'accès personnalisés ont été développés dans chaque catégorie.[Lechani et boughanem 2005].

4.3.1 Les systèmes de recommandation :

Les systèmes de recommandation se reposent sur les informations liées à la recherche de l'utilisateur en cours, dans le but de l'aider, soit en lui sélectionnant les liens pertinents dans une page, soit en lui recommandant des documents similaires au document en cours de lecture. Dans cette catégorie de systèmes, le profil de l'utilisateur obtenu centre son intérêt à court terme. Parmi les systèmes de recommandation: Letizia, WebACE, WebMat.

4.3.2 Les systèmes d'accès contextuel :

Dans les systèmes d'accès contextuel le profil utilisateur est construit prenant en considération le comportement de ce dernier qui peuvent être l'historique des interactions de l'utilisateur avec des applications Web. Le profil de l'utilisateur dans ce cas, reflète son centre d'intérêt à long terme. Parmi eux, on citera les systèmes Watson, Web Personae.

4.3.3 Les méta-moteurs de recherche personnalisée :

La quantité d'information sur le web couvert par un moteur, diminue avec l'accroissement de la taille du Web.Cela signifie que le besoin de l'utilisateur ne peut pas être satisfis en utilisant un seul moteur de recherche. A partir de là, les méta-moteurs de recherche sont apparus, qui accroissent la couverture de recherche en combinant les résultats issus de différents moteurs de recherche, comme MetaCrawler et DogPile.

L'inconvénient de ces méta-moteurs de recherche est qu'ils retournent une quantité considérable de résultats à l'utilisateur. C'est pourquoi, les méta-moteurs personnalisés sont apparus. Leur principe est d'intégrer le contexte et le centre d'intérêts d'utilisateur, pour lui permettre de retourner que les informations pertinentes, selon ses besoins. Nous pouvons citer dans ce cas le méta moteur *Inquirus* [Kechid,2009]

4.4 Approches de modélisation du profil utilisateur :

4.4.1 Définition :

On trouve dans le domaine de RI plusieurs définitions de la modélisation utilisateur, dont nous citons quelques-unes ci-dessous [Zemirli,2009] :

"*une connaissance à propos de l'utilisateur explicitement ou implicitement codée, utilisée par le système afin d'améliorer son interaction"*
"*User model is an explicit representation of the system of a particular user's characteristics that may be relevant for personalized interaction.*"

La modélisation des utilisateurs a fait l'objet de très nombreux travaux. Il s'avère que c'est là une problématique fort complexe qu'attestent les emprunts à différentes branches des sciences de l'information et des sciences de l'homme comme la psychologie, l'intelligence artificielle, la pédagogie, les interfaces Homme-Machine, etc. Les résultats restent mitigés avec indéniablement des échecs et des réussites.

4.4.2 Revue sur quelques approches de modélisation de l'utilisateur :

Les approches et techniques de la modélisation utilisateur peuvent être basées sur des modèles simples ou complexes dépendant de l'objectif final ou domaine d'application du système ; un effort de standardisation pour la généralisation de tels systèmes afin de produire des Shell a toutefois été mené et semble donner une meilleure portée au devenir des méthodes de modélisation de l'utilisateur.

Plusieurs techniques furent développées dans la littérature pour modéliser l'utilisateur, elles diffèrent cependant dans l'approche de représentation et construction du profil. En effet, le contenu informationnel du profil dépend fortement de l'application.

- Selon [Tamine,2007], tout système mettant en Œuvre des méthodes de modélisation de l'utilisateur inclut en partie les paquets d'informations suivants :
  - Des informations personnelles associées à l'utilisateur telles que l'âge, le pays, la langue,
  - Les préférences : peuvent être de différents niveaux telles que préférences de forme (style de la page, longueur d'un document) et préférences de domaine permettant de cibler le centre d'intérêts de l'utilisateur,
  - Historique de l'utilisateur : les interactions passées de l'utilisateur représentent une source pour prédire ses intentions et lui recommander des objets.

- Selon [Kechid,2009], Un profil peut être modélisé directement à partir des besoins et des préférences des utilisateurs ou découvertes par des méthodes de *data mining* à partir de leur comportement passé.

- Selon [Ilhem Boussaid,2005], pour la construction un profil utilisateur, donc, ils n'ont pas besoin d'une technique d'apprentissage de profil. Il y a trois principaux types de ces systèmes:
  - Systèmes qui acquièrent l'information du profil utilisateur d'une base de données. Par exemple, les systèmes de commerce électronique ([Amazon, 2001, qui extraient l'information à partir d'une base de données des produits et gardent une liste d'achat comme profil.
  - Systèmes de filtrage collaboratif qui gardent une matrice avec les estimations utilisateur/article sous forme de profil.
  - Les systèmes qui créent un profil initial à partir d'un stéréotype qui correspond à un condensé des caractéristiques les plus représentatives d'un groupe - ou classe - d'individus, que nous pouvons qualifier de valeurs par défaut. Des combinaisons de couples (items, valeurs) définissent différents stéréotypes telsque *novice*,

avancé, *expert*. L'utilisateur est donc associé à une des catégories élaborées et hérite de ses propriétés, et ne le modifient pas.

- Selon [Bradley 00], [Mobasher 00],le modèle de profils sont utilisés en général par les systèmes de recommandation. L'idée principale est de stocker des exemples d'objets que l'utilisateur a jugés comme pertinents ou comme inintéressants pour ensuite comparer le contenu des profils des différents utilisateurs entre eux. S'il y a des similarités entre deux profils le système va recommander à l'utilisateur des objets que d'autres clients de profils semblables ont appréciés. Comme dans certains cas les objets peuvent être volumineux (comme des livres, rapports etc.), dans ces cas il est possible de ne stocker que leurs identifiants.

- Selon [Bouhini,2014], le contexte informationnel social de l'utilisateur est composé du profil social de l'utilisateur $PS(u)$, construit à partir de ses activités d'annotation et de celui de son voisinage social (ses relations sociales) $PV(u)$, construit à partir des annotations des utilisateurs de son voisinage.

- Nous citons dans cette catégorie, les travaux des auteurs [V.Raghavan and H.Sever, 1995] qui représentent le profil utilisateur par son historique de recherche en se servant comme une base de données des requêtes soumises précédemment ainsi que leurs résultats associés. Cette base de données est utilisée pour sélectionner les requêtes les plus similaires à une requête en cours d'évaluation. Aussi l'approche proposée par l'article [B.Tan, X.Shen, and C.Zhai,2006], où le profil utilisateur est représenté par l'historique des requêtes passées et l'historique des clicks sur les résultats sélectionnés par l'utilisateur.

- Dans [Daoud,09], l'auteur divise la représentation ensembliste en trois sous-modèles de représentation :
  - Un ensemble de termes pondérés où chaque terme représente un centre d'intérêt possible de l'utilisateur,
  - Un vecteur de termes pondérés représentant un centre d'intérêt,

– Un ensemble des vecteurs de termes pondérés dont chacun représente un centre d'intérêt.

4.4.3 Approches d'intégrations de profile dans le processus d'accès a l'information :

L'intégration du profil utilisateur revient a l'utilisation de ce dernier dans l'une des principales étapes de l'évaluation de la requête : reformulation, calcule du score de pertinence de l'information ou présentation des résultats de recherche nous mentionnons quelques approches :

- [Sieg 2005]utilise une approche de personnalisation basée sur la reformulation de requêtes des utilisateurs ,cela sert a décrire une requête plus riche traduisant le contexte adéquat a la recherche en utilisant une variante de l'algorithme de Rachio .en effet ,le contexte de recherche est représenté par une paire de catégories de la hiérarchie de catégories de « yahoo »,la première représente la catégorie adéquate à la requête et similaire à l'une des centres d'intérêts de l'utilisateur et la deuxième représente la catégorie a exclure durant la recherche .

- [Tamine 2007]intègre le profile utilisateur dans la fonction d'appariement requête-document. Utilise les centre d'intérêts dans la fonction d'appariement du modèle de RI « le modèle bayésien ».la valeur de pertinence d'un document vis-à-vis une requête n'est plus fonction de la requête seule mais en plus du centre d'intérêts de l'utilisateur qui l'a soumise.

- [Vishnu Challem 2007]intègre le profile utilisateur dans la présentation des résultats, ils sont basés sur le ré-ordonnancement des résultats de recherche. Le ré-ordonnancement est basée sur la combinaison entre le rang initial du document et le rang résultant d'une mesure de similarité entre le document et le profil utilisateur.

- [Saulier *et al*., 14].Les systèmes de bases de données proposent des langages de requêtes avancés qui supportent la notion de préférence. Il en est ainsi pour les langages comme Preference SQL ou SQL /f ou pour les opérateurs de type « Best of

», et « Skyline ». La reformulation de la requête peut utiliser de nouveaux critères de sélection contenus dans le profil utilisateur. Le PU est composé d'un ensemble de prédicats pondérés. Le poids d'un prédicat exprime son intérêt relatif pour l'utilisateur. Il est spécifié par un nombre réel compris entre 0 et 1. La reformulation d'une requête à l'aide d'un profil consiste à ajouter dans la requête initiale les prédicats de sélection les plus pertinents,

5 Évaluation d'accès personnalisée àl'information :

L'évaluation des performances d'un modèle de RI, permet de paramétrer le modèle, d'estimer l'impact de chacune de ses caractéristiques. Dans la plupart des protocoles d'évaluation des systèmes d'accès à l'information, les mesures sont construites à partir des jugements exprimés par des utilisateurs ou par des experts. Ces mesures concernent plusieurs critères : le temps de réponse, la pertinence, la qualité et la présentation des résultats. [Zemirli,2009]

Les méthodes d'évaluation largement adoptées en recherche d'information, sont empiriques. Elles sont souvent basées sur une évaluation d'avantage quantitative que qualitative. En effet, les résultats obtenus sont issus de la comparaison de mesures et de métriques en termes de rappel et précision, sur les réponses fournies par le système relativement à celles issues des réponses attendues qui constituent le référentiel. Ce type d'évaluation, est orienté vers une approche comparative de plusieurs systèmes reposant sur le principe d'évaluation des collections de test. Bien qu'adopté par des campagnes d'évaluation de référence en recherche d'information, tels que TREC, il est cependant contesté notamment en raison de la non considération ni du contexte dans lequel se fait la recherche, ni de la perception de la pertinence des utilisateurs dans ce même contexte. L'introduction de la dimension utilisateur dans le processus d'accès à l'information, accentue d'avantage la difficulté d'évaluation. En effet, en raison du caractère subjectif des utilisateurs d'une part et du caractère dynamique ou adaptatif du système, d'autre part, il est difficile de fournir des valeurs absolues aux métriques. On cite dans ce qui suit les principaux problèmes liés à l'évaluation des systèmes personnalisés puis présentons quelques recommandations qui augmentent la fiabilité des résultats qui en sont issus.

5.1 Problèmes de l'évaluation :

La rationalité des résultats issus d'un scénario d'évaluation d'un système personnalisé est compromise pour les principales raisons suivantes [Chi, 2001]:

- Si l'évaluation des variables indépendantes est effectuée par différents utilisateurs, alors des différences personnelles exogènes à l'expérimentation, telles que l'intelligence, la capacité de raisonnement, l'expérience etc.., ont un impact sur les valeurs des autres variables.

- Si le même utilisateur est impliqué dans des différents scénarios d'évaluation, alors son expérience passée avec le système peut influer sur sa perception de la pertinence.

- Les conditions de déroulement des expérimentations ont un impact non négligeable sur les résultats : configuration des machines (processus lent conduit à une lassitude de l'utilisateur), interfaces peu ergonomes, lieu inapproprié (bruit, présence de caméras) etc…

- La validation interne est difficile à mettre en Œuvre. En effet, en raison des problèmes évoqués ci-dessus, il n'est pas aisé de séparer l'effet intrinsèque des variables pertinentes du système indépendamment des facteurs exogènes. La validation externe est dès lors difficile également. Elle l'est d'autant plus que la généralisation des résultats obtenus lors de la validation interne doit considérer une combinaison des divers types d'utilisateurs, des variables d'expérimentation et des situations d'utilisation.

5.2 Recommandations :

Dans le but d'éviter des erreurs de mesures dus à l'aspect subjectif de l'évaluation, on cite dans ce qui suit quelques recommandations [Chi 2001] :

- Définir un nombre suffisant de groupes d'utilisateurs avec des effectifs adéquats,

- Isoler au mieux les utilisateurs,
- S‘assurer de l‘ergonomie des applications,
- Préparer un canevas unique qui décrit le protocole d‘expérimentation et l‘adresser à l‘ensemble des utilisateurs,
- Effectuer des expérimentations pilotes avant de passer aux expérimentations effectives,
- Les utilisateurs ne doivent pas être informés des facteurs à évaluer dans le système (ex avec et sans modèle utilisateur) ; leur appréciation doit porter sur des aspects perceptibles, liés à son fonctionnement,
- Les variables exogènes (âge, expérience, aptitudes …) doivent être identifiées explicitement et leur influence mesurée pour être prise en compte dans le processus global d‘évaluation.

6 Contribution recherche sémantique et personnalisée :

6.1 Représentation sémantique du profil utilisateur :

Parmi les *caractéristiques* de l’utilisateur, les plus populaires sont ses connaissances, intérêts, objectifs, contexte, et traits individuels. Comme *techniquesde représentation*, parmi les plus populaires dans les systèmes adaptifs hypermédia on peut mentionner la représentation basée sur des *mots* clefs, sur une *réseau sémantique* (les mots clefs qui comportent une liaison sémantique sont groupés, comme par exemple dans Wordnet), ou la représentation basée sur des *concepts* : des poids sont associés aux concepts des réseaux sémantiques existants, comme par exemple des taxonomies, TopicMap ou, plus récemment, les ontologies.

Parmi les approches qui utilisent des ontologies afin de développer le profil utilisateur par rapport à ses différentes caractéristiques, on peut mentionner :

- Le système *CommonCV* [Trichet *et al.*, 2003], focalisé sur ses connaissances, modélisées et gérées par rapport à une taxonomie de compétences,

- Le système *Bibster* [Haase *et al.*, 2005], dans lequel le profil utilisateur est constitué comme un ensemble pondéré de concepts sélectionnés de la taxonomie ACM qui expriment les intérêts à long terme de l'utilisateur. Ce profil est utilisé pour fournir à l'utilisateur des réponses personnalisées à ses requêtes.

- Le système *Foxtrot* [Middleton *et al.*, 2004] qui développe le profil utilisateur par rapport à son activité courante (ses objectifs) en associant une valeur d'intérêt pour chaque concept de la même taxonomie ACM, afin de recommander à l'utilisateur des documents pertinents pour son activité.

- [Farzan *et al.*, 2006], Dans le but de recommander des cours conformément aux intérêts à long terme des utilisateurs, le système *CourseAgent* adopte une même approche, mais en utilisant une taxonomie des profils d'emplois

- D'après[Douad,09] on a deux approches :

➢ **Le profil comme une liste de concepts :**

Nous proposons un modèle de représentation préliminaire du profil de l'utilisateur défini au niveau i de l'ontologie par une liste de concepts, noté Pu = {C 1 i , C2 i , . . . , Cn i }. Notre première intuition derrière ce modèle de représentation consiste à représenter le profil utilisateur par un ensemble de concepts ayant un niveau de spécificité moyen de l'ontologie de l'ODP et à la fois efficace dans le sens où il permet d'améliorer la précision de recherche des requêtes corrélées à un même centre d'intérêt. Bien que ce modèle de représentation du profil permet d'avoir un niveau de personnalisation efficace au niveau de la session, nous proposons d'enrichir cette représentation dans le but de recouvrir deux aspects liés principalement à la corrélation des concepts du profil et à sa spécificité.

➢ **Le profil comme un graphe de concepts :**

Le but fondamental de la représentation du profil de l'utilisateur sous forme d'un graphe est de décrire le profil sous forme d'un ensemble de concepts sémantiquement liés à travers l'ontologie de l'ODP et comprenant des concepts de plusieurs niveaux spécifiques. Formellement, la structure du graphe G=(V,E) est constituée d'une composante hiérarchique formée par les liens de type "is-a" et une composante non hiérarchique formée par des liens de différents types prédéfinis dans l'ontologie de l'ODP, où : – V est un ensemble de nœuds pondérés, représentant les concepts d'intérêts de l'utilisateur, – E est un ensemble d'arcs entre les nœuds du graphe V , partitionné en trois sous-ensembles T, S et R, tel que : – T correspond à la composante hiérarchique du profil utilisateur contenant les liens de type "is-a"

6.2 Pondération sémantique des termes

Dans [Zargayouna'04], le poids sémantique $SemW$ du terme t au sein d'une balise b d'un document d correspond à la somme de son poids TF-ITDF et des poids des termes ti qui lui sont proches sémantiquement. Les termes ti considérés comme étant sémantiquement proches sont ceux dont la similarité est supérieure à un seuil initialisé en fonction de la similarité entre le terme t et la balise b.

$$SemW(t,b,d) = TFITDF(t,b,d) + \frac{\sum_{i=1}^{n} Sim_{ZS}(t,t_i) * TFITDF(t_i,b,d)}{n}$$

L'estimation de ce seuil peut toutefois poser problème dans le cas où les balises d'un document ne comportent pas de sémantique propre, par exemple si elles représentent la structure logique d'un document. De plus, le TF-ITDF se basant sur le nombre d'occurrences d'un terme dans un type de balise précis semble moins pertinent dans un environnement dans lequel la structure des documents est fortement hétérogène, et dans lequel les balises rencontrées sont très souvent différentes.

7 CONCLUSION :

Nous avons présenté au cours de ce chapitre les principales notions et concepts de la recherche d'information personnalisée (RIP) telle que la notion de contexte, profile utilisateur et la Pertinence contextuelle. Nous avons aussi cité les déférents services de personnalisation à savoir Le filtrage des résultats, Le re-ordonnancement des résultats de la requête et La recommandation. Ensuite, nous avons étudié l'Architecture générale d'un système de recherche d'information (SRIP) ou on a dégagé deux fonctions fondamentales qui sont la gestion des profils et la Sélection des informations ainsi que Les prototypes des systèmes d'accès personnalisés et les Approches de modélisation du profil utilisateur. Et nous avons étudié l'évaluation des systèmes d'accès personnalisé à l'information où nous avons citée es Problèmes de l'évaluation et quelques recommandations. En fin nous avons accéder à la contribution des deux recherche sémantique et personnalisé en situant des approches sur la pondération des termes et des représentations sémantiques de profile utilisateur.

Chapitre 3 :

Reformulation de la requête

Chapitre 3 : Reformulation de la requête

1. INTRODUCTION :

Les systèmes de recherche d'information (SRI) sont des systèmes capables de présenter un ensemble de documents qui répondent aux besoins de l'utilisateur. Le processus de recherche d'information se base essentiellement sur la mise en relation des documents du corpus et des besoins en information de l'utilisateur. Ces besoins sont structurés et formulés sous forme de requêtes. Les documents sélectionnés et retournés à l'utilisateur sont ceux jugés pertinents par rapport à ses besoins. La notion de pertinence est déterminée selon plusieurs critères, qui dépendent du SRI et de l'utilisateur.

Les requêtes des utilisateurs sont reformulées parce qu'elles expriment le besoin d'information d'une manière pas assez satisfaisante pour un système de recherche d'information. Nous pouvons distinguer deux causes principales à la mauvaise qualité des requêtes des utilisateurs du point de vue du système : le choix des termes et celui des paramètres. Le premier problème est que l'utilisateur n'a pas forcément une idée précise de la collection de documents sur laquelle il exécute sa requête (surtout sur le Web). Il n'a donc pas la capacité de prédire les termes à la fois correspondants à son besoin d'information et apparaissant dans les documents pertinents. Deuxièmement, pour des raisons comme le manque d'expertise ou de temps, Une approche de reformulation de la requête est donc souvent nécessaire pour que les requêtes des utilisateurs profitent au mieux des capacités fournies par un système de recherche d'information [Bissan Audeh, 2014].

Plusieurs définitions de reformulation de requête ont été données dans la littérature, nous en présentons quelques-unes :

- La reformulation de requête est proposée comme une méthode élaborée pour larecherche 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. [Lynda Tamine,2000]

- La reformulation de la requête est la procédure qui permet d'organiser les termes des requêtes dans une structure autre que le sac de mots, en utilisant des fonctions, comme la proximité et la pondération, fournis par le modèle de recherche

Dans ce chapitre on va commencer tout d'abord de présenter la reformulation des requêtes dans la recherche d'information classique on citons ces trois types (manuelle ,automatique et semi-automatique) ensuite on va définie la reformulation sémantique des requêtes on mentionnant ces déférents ressources externes avec quelques approches après on va expliquer l'intérêt de la reformulation personnalisée des requêtes et comment exploiter l'historique de recherche de l'utilisateur pour reformuler la requête on donnant quelques approches et on terminant par quelques défauts des approches de reformulation de requête.

2. La reformulation classique :

Le but essentiel d'un système de recherche d'information est de permettre à l'utilisateur d'avoir un résultat satisfaisant (des documents pertinents) par apport à son besoin exprimé par une requête. La possibilité de reformuler la requête initiale s'avère intéressante dans le processus de la RI. Cela fera en sorte que le résultat retourné soit plus pertinent. Il existe trois méthodes de reformulation de requêtes :

2.1 La reformulation manuelle :

Cette approche est associée aux systèmes de recherche booléens Elle consiste à présenter à l'utilisateur une liste de documents jugés pertinents en réponse à la requête initiale. C'est à l'utilisateur de sélectionner à partir des documents pertinents ceux dont lesquels le système va extraire les termes à rajouter à la requête initiale dans le but d'effectuer une nouvelle recherche. [Merad&Asnoun, 09]

2.1 La reformulation semi-automatique :

Cette technique nécessite l'intervention de t'utilisateur qui doit identifier et sélectionner les documents pertinents et les documents non pertinents. [Caro, 97]

2.2 La reformulation automatique :

Lorsque le feedback de pertinence s'accompagne d'une adjonction (et/ou) suppression de termes, on parle de reformulation automatique. La requête de l'utilisateur est remaniée automatiquement, pour intégrer les descripteurs des documents jugés pertinents ou rejetés. On trouve différentes variantes de cette technique : celles qui sont utilisées automatiquement pour reformuler la requête en augmentant le poids des termes présents dans les documents jugés pertinents et inversement pour diminuer les poids des termes jugés non pertinents.

Le problème avec la re- formulation automatique est l'estimation des « bons » termes qui peuvent conduire effectivement à une amélioration du processus de recherche car l'introduction des termes inappropriés peut entraîner un silence ou au contraire augmenter un bruit.

3. La reformulation sémantique :

Très tôt dans le domaine de la recherche d'information, il a été clair que la ressemblance syntaxique entre les termes d'une requête et les termes d'un document sur laquelle se base l'évaluation de la pertinence n'était pas suffisante pour répondre au besoin d'information de l'utilisateur. Ce problème a été cité pour la première fois en 1988 en parlant des études réalisées par John O'Connor sur la relation entre la fréquence des mots et l'indexation de documents, autour des années 1960. Ces études ont montré que parmi 23 documents manuellement indexés sous le mot « toxicity », il n'en existe que 11 contenant le terme « toxicity », alors que plus de la moitié de ces documents réellement pertinents ne contient pas de mots dont l'orthographe est liée au lemme « toxicity ». La littérature de la recherche d'information contient un grand nombre d'études qui ont essayé d'intégrer la sémantique dans un SRI.

Un système de recherche d'information peut impliquer la sémantique au niveau de l'index ou de la requête. Si la sémantique est intégrée dès la phase d'indexation, toute la procédure de la recherche (la représentation des documents et des requêtes, la mise en correspondance, et éventuellement la reformulation de la requête) sera faite en utilisant des concepts au lieu de mots. Dans le deuxième cas, l'intégration de la sémantique au niveau de la requête, la notion sémantique est utilisée au moment de l'interrogation ; le modèle de recherche par mots-clés n'est donc pas affecté. La couche sémantique dans ce cas sert à identifier les concepts de la requête et à choisir les termes qui les représentent le mieux afin de générer une nouvelle requête textuelle de « meilleure qualité ».

Bien qu'un index sémantique signifie une requête sémantique, il existe des cas où l'on souhaite garder un index par mots-clés, mais en introduisant la sémantique au niveau de la requête. Dans ces occurrences, la requête reformulée est constituée de mots, et le rôle de la sémantique est de bien choisir ces mots.

3.1 Les Ressources sémantiques :

Durant de longues années pendant lesquelles les chercheurs ont essayé d'appliquer la sémantique à divers domaines, plusieurs noms ont été utilisés pour désigner les ressources sémantiques (ontologie, thésaurus, dictionnaire, terminologie, taxonomie, etc.). La signification de chacun de ces mots n'est pas là même selon les études ou les domaines dans lesquels ils sont employés. Il est donc préférable que chaque travail sur la sémantique clarifie précisément son intention dans l'utilisation de ces termes.

3.1.1 Les dictionnaires :

C'est une ressource qui contient les définitions des termes. Même si de nombreux dictionnaires proposent, en plus de la définition, des synonymes (et éventuellement des antonymes) pour un vocable donné, la structuration de ces ressources y reste peu développée.

3.1.2 Les taxonomies :

C'est une structure qui permet de contrôler le vocabulaire par un seul type de relation donnant la possibilité de généraliser ou de préciser un sens. Cette ressource peut être considérée comme une catégorisation pour un domaine précis. La figure 3.1 constitue un exemple d'une taxonomie des animaux où la relation exprimée est « est un(e) » /« comprend » selon la direction de la relation.

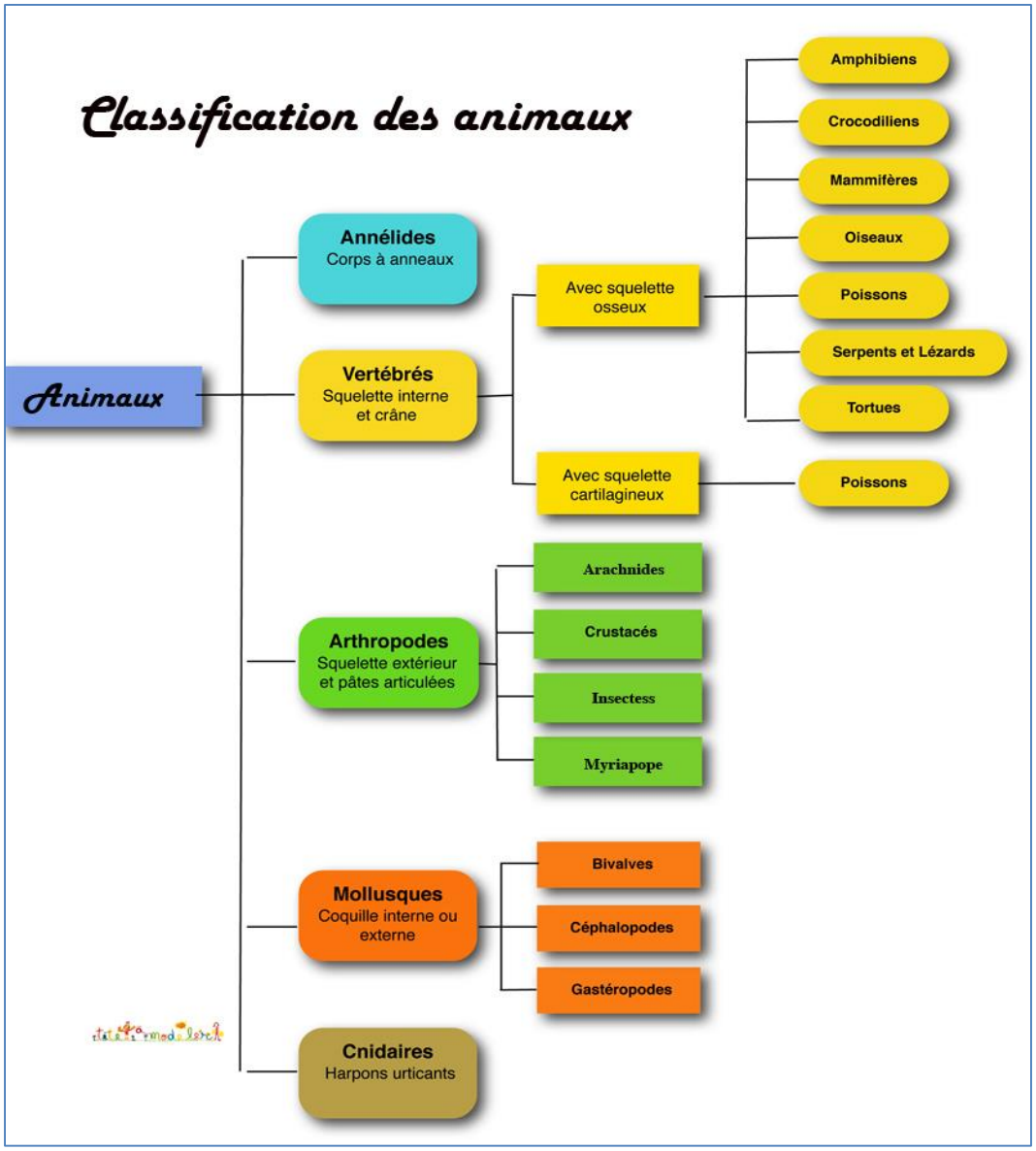


Figure 3.1: Exemple d'une taxonomie des animaux

3.1.3 Les thésaurus :

Un thésaurus est une ressource qui représente le vocabulaire par des concepts. Chaque concept a un identifiant, une définition, et contient les termes qui peuvent être employés pour le dénoter. Ainsi, il peut être considéré comme une extension d'un dictionnaire, car il contient une définition des termes qui appartiennent au même concept, par contre les relations entre ces concepts sont plus riches et plus formalisées par rapport à un dictionnaire. Contrairement à une taxonomie, le thésaurus peut avoir une structure non hiérarchique entre ces éléments. Les relations sémantiques les plus fréquentes dans un thésaurus sont les relations d'équivalence (synonymie/antonymie) et les relations hiérarchiques (hyperonyme/hyponyme).

Wordnet est l'exemple le plus connu d'un thésaurus, selon notre définition, même si beaucoup d'études le considèrent comme une ontologie ou une taxonomie lexicale. Dans tous les cas, Wordnet est la ressource lexicale la plus célèbre de la langue anglaise. La figure 3.2présente une visualisation graphique du mot « book » dans WordNet, où les concepts sont appelés « synsets ». Chaque synset contient des termes qui possèdent des relations de synonymie entre eux. Les concepts sont liés par d'autres types de relations sémantiques (antonymie, hyperonymie, etc.). Un mot peut appartenir à un concept ou à plusieurs s'il est polysémique.

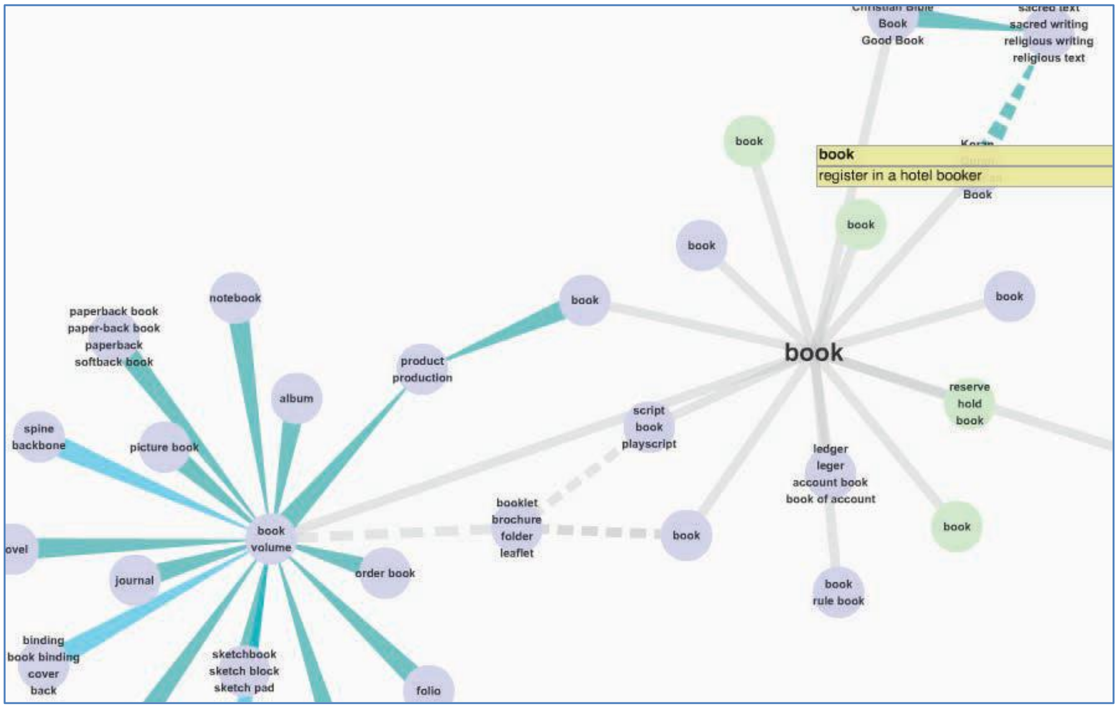


Figure 3.2:Le mot « book » dans le thésaurus anglais WordNet

3.1.4 Les ontologies :

Une ontologie est une collection de concepts bien définis qui décrivent un domaine spécifique. Les relations entre ces concepts peuvent être différentes d'une ontologie à une autre. Par exemple, les relations existantes dans une ontologie du domaine juridique n'ont pas les mêmes significations que celles d'une ontologie de la génétique. Chaque concept dans une ontologie peut avoir une définition et des propriétés qui le caractérisent. Contrairement à un thésaurus, nous pouvons y trouver des concepts abstraits, c'est-à-dire qui ne sont pas exprimable par un terme du langage humain, mais qui peuvent être compréhensibles par leurs liens avec d'autres concepts. Compte tenu de la variété des relations qui peut être exprimée dans une ontologie, la structure hiérarchique telle qu'on la trouve dans une taxonomie n'est pas forcément garantie. Avec le niveau élevé de modélisation fourni par les ontologies, des langages de représentation et de spécification ont été proposés pour simplifier la manipulation de ces ressources. Les langages les plus connus sont issus du W3C, comme RDF (Resource Description Framework), RDF Schema, OWL (Ontologie Web Language) et SPARQL.

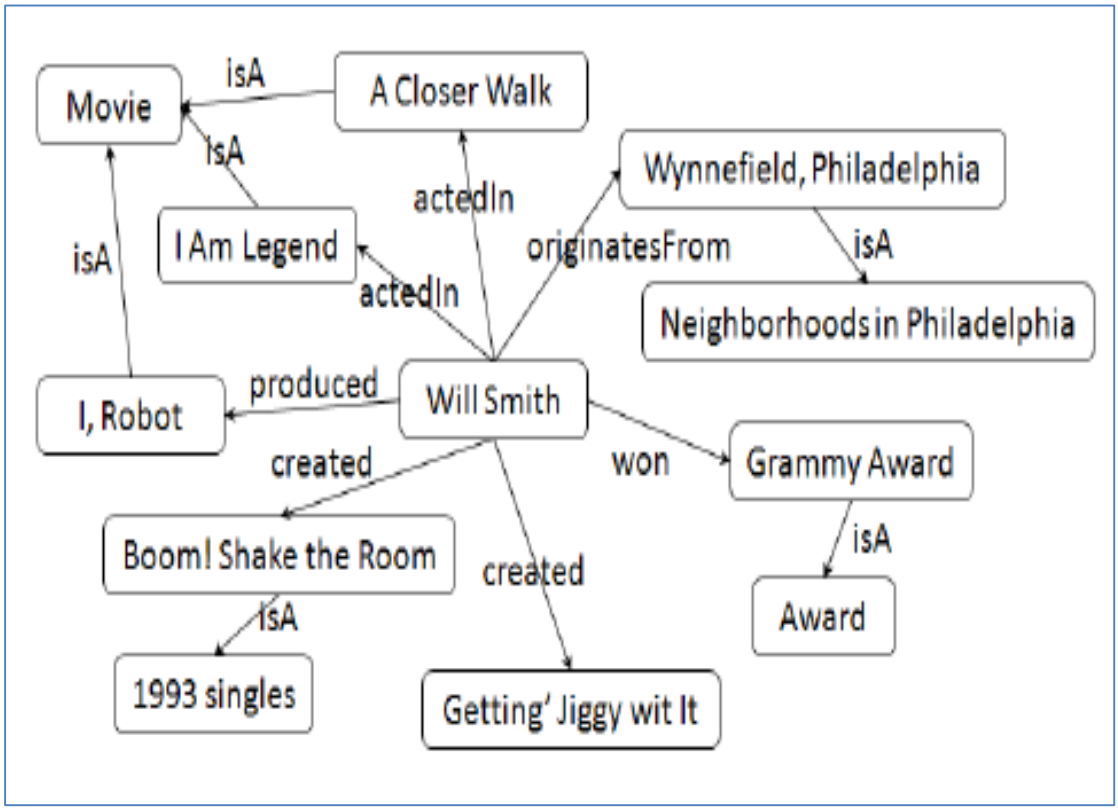


Figure 3.3:Un extrait de l'ontologie

3.2 Quelques approches de reformulation sémantique :

L'utilisation d'une ressource sémantique pour l'expansion de la requête permet d'éviter la complexité élevée des approches basées sur toute la collection de documents et les limites de celles fondées sur le principe de retour de pertinence comme la dépendance de plusieurs paramètres et de la qualité du petit nombre de documents de pseudo retour de pertinence. De plus, la requête étendue avec ces techniques est plus compréhensible par les humains que les méthodes statistiques, car ces dernières peuvent ajouter des mots qui n'ont pas de lien logique (au sens humain) avec les termes de la requête originale.

- En 2003 Navgli propose une approche qui consiste à extraire des concepts et relations d'une ontologie et les ajouter dans la requête initiale en se basant sur les liens entre les termes

- NavigliandVelardi ont proposé une étude sur les meilleurs termes d'expansionà choisir une fois la désambiguïsation (le choix du synset dans Wordnet) faite. Alorsque cette étude mentionne que l'utilisation de la définition (le glossaire de Word-Net) est plus performante pour l'expansion de la requête que les synonymes et leshyperonymes, les résultats ne sont pas concluants à cause des limites des expériences. Avigliand Velardi[2003]

- Cronen-Townsend et al propose une approche pour pondérer les termes de la requête par les scoresde clarté. Dans cette dernière, les termes qui ont un score élevé sont considéréscomme précis et ne sont donc pas étendus, ceux avec un score faible sontconsidérés comme ambigus et ne sont pas étendus non plus. Uniquement les termesde clarté « moyenne » sont étendus par des synonymes de WordNet. La désambiguïsation pour trouver ces synonymes et la méthode pour les ajouter à la requête ne sont pas clairs dans ce papier, de plus les résultats montrent que leur approche n'est pas très efficace sur la précision moyenne mais elle est plus adaptée à un contexte question/réponse.[Cronen-Townsend et al. 2002].

- En 2009 A. Bouramoul et al base sur la prise en compte du contexte utilisateur via son profil, elle est utilisable directement sans phase d'analyse ou d'apprentissage, ces dernières n'utilisent que les relations d'équivalence et de subsomption et n'exploite pas toutes les relations sémantiques offertes par une ontologie. En rajoutant des termes issus de son contexte de recherche actuelle, ces termes sont extraits de la base des contextes utilisateurs. Une fois la requête reformulée elle sera envoyée au module de recherche qui prend en charge la délivrance des résultats à l'utilisateur. [A. Bouramoul.2009]

4. La reformulation personnalisée :

Le but fondamental de la reformulation de requêtes par utilisation de profil consiste à cibler la recherche des documents pertinents par enrichissement de la requête par des

termes issus du profil utilisateur dans le but de mieux répondre au besoin en information de l'utilisateur.

4.1 Exploiter le profil d'utilisateur pour la reformulation :

Pour améliorer une requête, cette méthode utilise les informations collectées de façon indirecte à la suite des interactions des utilisateurs dans un system de recherche d'information. Le plus fréquent est d'utiliser les historiques des sessions de recherche générés par l'ensemble des utilisateurs, ce qui demande un volume important de données. Pour cette raison, nous trouvons ce genre d'approche dans un contexte de recherche sur le Web.

Lors de la saisie de sa requête, l'utilisateur formule son intérêt dans son langage naturel. En effet, pour dire qu'il veut avoir des documents ou des liens concernant l'hôtellerie de luxe à un prix raisonnable, l'utilisateur aurait tendance à écrire la requête suivante : "hôtel de luxe et pas trop cher". Cette requête est certes claire mais présente des termes de liaison qui pourraient être enlevés : et, de. Reformuler une requête, dans ce contexte, a donc pour objectif de coordonner le langage de recherche naturel (utilisateur) et le langage d'indexation des systèmes de recherche d'information. Cela consiste à générer une nouvelle requête en utilisant d'autres données. Pour la personnalisation, les données utilisées sont principalement relatives à l'utilisateur. Il s'agit alors d'un enrichissement de la requête. Dans ce cadre, les données intégrées sont les éléments des centres d'intérêt ainsi que des préférences. [Myriam Hadjouni Krirn,2013]

4.2 Quelques approches de reformulation personnalisée :

- Cui et al Leur idée est de profiter du lien entre les termes des requêtes et les documents choisis dans les sessions de recherche des utilisateurs. Si un document est souvent ouvert pour la même requête dans plusieurs sessions de recherche, les termes de ce document sont probablement pertinents vis à vis des termes de la requête, et sont donc de bons candidats pour l'expansion de nouvelles requêtes similaires.[Cui et al. 2003]

- Wang and Zhaiproposent l'extraction de motifs (patterns) à partir de ces historiques. Ces motifs sont utilisés pour décider si un terme associé à la requête doit être ou non ajouté.[Wang and Zhai 2008]

- Dans le but d'augmenter la pertinence des données semi-structurées et d'offrir un accès personnalisé a ces derniers, en 2006 Zayani a proposé une architecture d'un système d'adaptation avec des fonctionnalités permettant l'enrichissementd'une requête à partir du profil utilisateur et de le mettre à jour enfonction des requêtes de l'utilisateur. Les auteurs proposent un modèle utilisateuravec des caractéristiques permanentes et évolutives pour stocker lesinteractions. Les informations permettent l'évaluation de la requête afin de L'enrichir et rendre le résultat plus pertinent. Les requêtes sont enrichies avec des termes représentant les intérêts de l'utilisateur.

- En 2004 Le système AIRA (Adaptive Information Research Assistant) Construit un profil utilisateur afin d'enrichir la requête. Lors de la recherche, le système identifié un contexte à partir de la requête et du profil de l'utilisateur. L'enrichissement de la requête passe par deux phases :
  1) Apprentissage du vocabulaire
  2) Apprentissage de requêtes.

  Le profil de l'utilisateur est mis à jour en fonction de sa recherche. A l'issue des deux phases les documents les plus pertinents sont retournés à l'utilisateur. Le fonctionnement du système repose sur l'hypothèse que l'utilisateur maintient à jour un ensemble de documents ou de références qui représentent son profil et intérêts qui sont représentatives de ses centres d'intérêts. C'est à partir de cet ensemble que le système extrait et structure les ressources représentatives de l'utilisateur. L'apprentissage du vocabulaire vise à identifier les meilleurs termes pour la recherche de documents et l'apprentissage de la requête recherche la combinaison optimale des termes sélectionnés dans la première phase. Lors de cette phase les requêtes sont soumises aux moteurs de recherche et les documents réponses sont évaluées et les requêtes pondérées en fonction de l'évaluation. Pour chaque requête sélectionnée de nouveaux termes sont ajoutés.

5. Les defaults des approches de la reformulation des requêtes :

Le fait de reformuler automatiquement la requête de l'utilisateur introduit plusieurs problèmes quel que soit l'approche utilisée pour l'expansion, nous citons le risque de dérive de la requête et la qualité des termes d'expansion.

La dérive de la requête signifie l'altération du but de la recherché C'est un phénomène qui se produit lors de l'expansion (généralement automatique) de la requête. Les nouveaux termes ajoutés à une requête peuvent pousser les résultats vers un sens différent de celui cherché à la base. La dérive de la requête est souvent un comportement « mauvais » que l'on souhaite éviter.

Pour la plupart des approches d'expansion de la requête, l'ajout de nouveaux termes est systématique pour toutes les requêtes, même si beaucoup de ces approches considèrent l'expansion sélective comme une étape importante dans leurs perspectives. Nous trouvons dans la littérature des méthodes qui mesurent la qualité d'une requête ou celle d'un terme de la requête. Par exemple, un peu d'étude propose une métrique de clarté, basée sur le calcul de l'entropie entre le modèle de la requête et celui de la collection.

Un défi majeur pour les approches locales est le choix des paramètres, surtout le nombre de documents de retour de pertinence. Plusieurs études se sont intéressées à cette question sans vraiment donner de réponse. Néanmoins, ces études ont confirmé le souci de stabilité, où un bon réglage des paramètres ne marchera pas forcément sur d'autres collections de tests, et même pas sur deux requêtes différentes sur la même collection de documents.

6. CONCLUSION :

Au cours de ce chapitre nous avons présenté d'abord la reformulation des requêtes dans la recherche d'information classique ou on a cité ces trois types (manuelle ,automatique et semi-automatique) , ensuite on a défini la reformulation sémantique des requêtes ou on a mentionné les déférents ressources externes avec quelques approches, après nous avons expliqué l'intérêt de la reformulation personnalisée des requêtes et comment exploiter l'historique de recherche de l'utilisateur pour reformuler la requête et on a donné quelques approches d'exemples, et en fin on a terminé par quelques défauts des approches de reformulation de requête.

Chapitre 4 :

Reformulation sémantique et personnalisée de la requête d'utilisateur

Chapitre 4 :reformulation sémantique et personnalisée de la requête d'utilisateur

1. INTRODUCTION :

Les approches en RI se sont orientées vers une nouvelle génération de systèmes de recherche basés sur l'accès personnalisée et sémantique à l'information. Le domaine de la RI contextuelle est apparu récemment comme une priorité, son objectif est d'exploiter le profil de l'utilisateur ainsi que les connaissances liées à la requête dans le but de répondre mieux aux besoins en information de l'utilisateur. De même l'émergence du Web sémantique a favorisé les travaux de recherche dans le domaine de la RI sémantique.
Afin de profiter à la fois des avantages de reformulation sémantique et personnalisée des requêtes on a proposé dans ce mémoire de combiner ces deux types. Dans ce contexte, nous proposons une approche de reformulation personnalisée basée sémantique des requêtes utilisateurs, cette reformulation est qualifiée de contextuelle car elle prend en compte la notion du contexte via les profils utilisateur pour modifier leur requête initiale. L'ajout de la sémantique lors de la recherche personnalisée vise à augmenter l'efficacité des SRI en améliorant leur pertinence et en permettant ainsi de prendre en compte les caractéristiques personnelles de l'utilisateur, ses intérêts, ses performances et l'historique de ses interactions avec le system de recherche pour exploiter les termes pertinents en utilisant l'ontologie WordNet. Ces éléments sont capitalisés dans notre système sous forme du partie statique et dynamique et sémantique pour être utilisé par la suite de reformulation sémantique personnalisée.

2. L'ontologie utilisé :

Le choix de la ressource sémantique à utiliser pour l'expansion de la requête est crucial et doit être en harmonie avec le domaine de la collection de documents recherchée. Quand nous sommes dans un contexte de recherche Web où le domaine de recherche de la requête ne peut pas être identifié par avance, une ressource sémantique non liée à un domaine spécifique est la mieux adaptée. Le choix le plus naturel dans ce cas est l'utilisation d'un thésaurus comme Wordnet.

2.1 WordNet :

WordNet est un réseau lexical électronique développé depuis 1985 à l'université de Princeton par une équipe de psycholinguistes et de linguistes du laboratoire des sciences cognitives. L'avantage de WordNet réside dans la diversité des informations qu'elle contient (grande couverture de la langue anglaise, définition de chacun des sens, ensembles de synonymes, diverses relations sémantiques). En outre, WordNet est librement et gratuitement utilisable. WordNet couvre la majorité des noms, verbes, adjectifs et adverbes de la langue Anglaise, qu'elle structure en un réseau de nœuds et de liens. Les nœuds sont constitués par des ensembles de termes synonymes (appelés *Synset*) Chaque synset est identifié par un numéro, et possède une définition (glossaire) qui décrit le concept port é par ce synset en utilisant des commentaires et/ou des exemples de son usage. Un terme peut être un mot simple ou une collocation (i.e. deux mots ou plusieurs mots reliés par des soulignés pour constituer le mot composé). Un exemple des synsets correspondant au mot « mouse » est donné dans la figure 4.1.

Noun

- S: (n) **mouse** (any of numerous small rodents typically resembling diminutive rats

having pointed snouts and small ears on elongated bodies with slender usually hairless tails)
- S: (n) shiner, black eye, **mouse** (a swollen bruise caused by a blow to the eye)
- S: (n) **mouse** (person who is quiet or timid)
- S: (n) mouse, computer mouse (a hand-operated electronic device that controls the

coordinates of a cursor on your computer screen as you move it around on a pad; on the bottom of the device is a ball that rolls on the surface of the pad) *"a mouse takes much more room than a trackball"*

Verb

- S: (v) sneak, **mouse**, creep, pussyfoot (to go stealthily or furtively) *"..stead of sneakingaround spying on the neighbor's house*"

Figure 4.1 : les concepts de Wordnet correspondants au mot mouse

Les concepts de WordNet sont reliés par des relations sémantiques. La relation de base entre les termes d'un même synset est la synonymie. Les différents synsets sont autrement liés par diverses relations sémantiques telles que la subsomption ou la relation d'hyponymie hyperonymie, et la relation de composition méronymie holonymie. Ces relations sont formellement définies comme suit :

1. La relation taxonomique (ou relation de subsomption), dite relationD'Hyperonymie / Hyponymie : X est un hyponyme de Y si X est un type de (kind of /Is a) Y. Y est alors dit hyperonyme de X. *Exemple :* {canine} a pour hyponymes {Wolf, Wild dog, dog}.

2. La relation d'Holonymie et son inverse la Méronymie :X est un méronyme de Y si X est une partie constituante (part of), substance de(Substance of) ou membre (member of) de Y. Y est alors dit un homonyme de X.*Exemple :* {car} a pour méronymes {wheel, engine, …}

Enfin, la figure 4.2, donne un exemple de sous-hiérarchie de WordNet correspondant au concept « dog ».

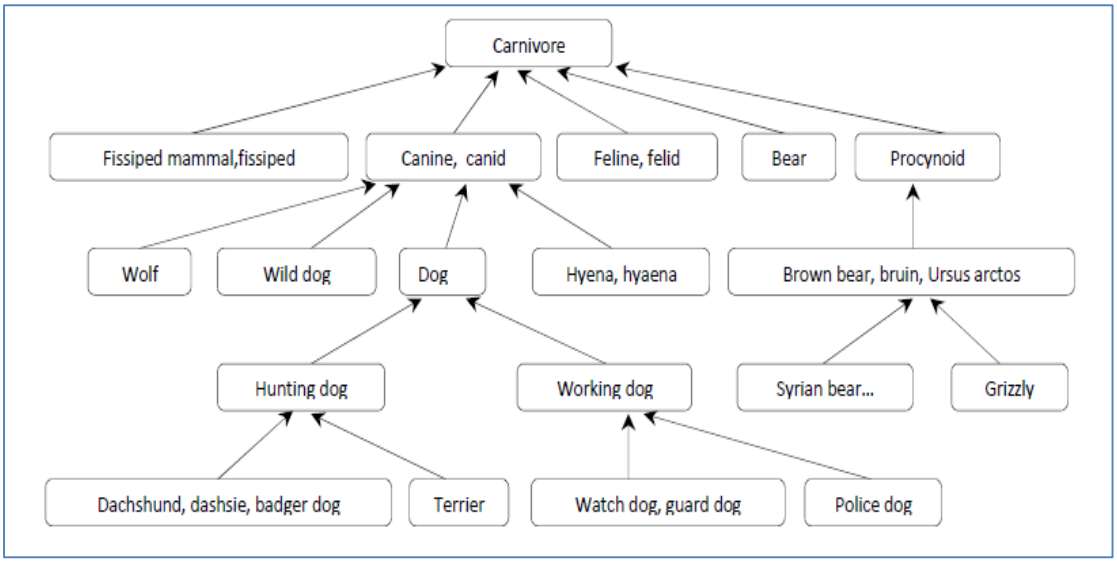


Figure 4.2 :Sous hiérarchie de WordNet correspondant au concept "dog"

3. Le profil utilisateur proposé :

Un profil utilisateur est défini comme « une source de connaissance qui contient des acquisitions sur tous les aspects de l'utilisateur qui peuvent être utiles pour le comportement du système »

L'utilisateur intervient en partie dans la définition de son profil, l'implication est donc directe. A la fin de chaque session de recherche, le système récupère automatiquement des informations qu'on suppose pertinentes pour enrichir le contexte.

Pour améliorer la représentation de l'utilisateur et ses intérêts, nous renforçons ce profil par l'introduction d'une sémantique ontologique dans laquelle les liens sémantiques sont fournis par une ontologie.

Pour rendre l'utilisation des profils utile pour la reformulation contextuelle des requêtes et utilisable dans un système de recherche d'information, nous proposons nos choix de profil qui se divise en trois partie :

Partie statique : elle contient les caractéristiques d'un profil d'identification qui se caractérise par une implication directe de l'utilisateur

Partie dynamique : elle contient les documents pertinents avec leurs termes pertinents à partir de l'interaction de l'utilisateur avec le système.

Partie sémantique : la partie qui sauvegarde les termes pertinents des documents pertinents, ainsi que pour sauvegarder les similarités de ces termes avec les concepts fournis par une ressource sémantique.

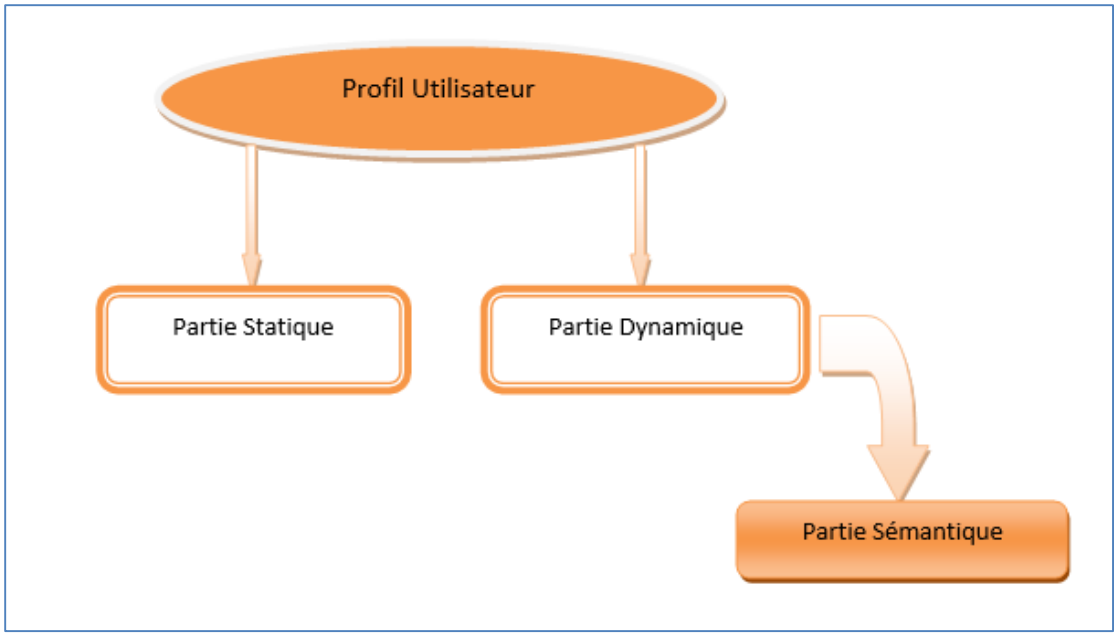


Figure 4.3: présentation de profil Utilisateur

3.1 Partie statique :

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 ...
- *Les intérêts et préférences :* domaine, domaine secondaire, spécialité…
- *Les compétences et niveau du savoir-faire :* profession, niveau d'étude...

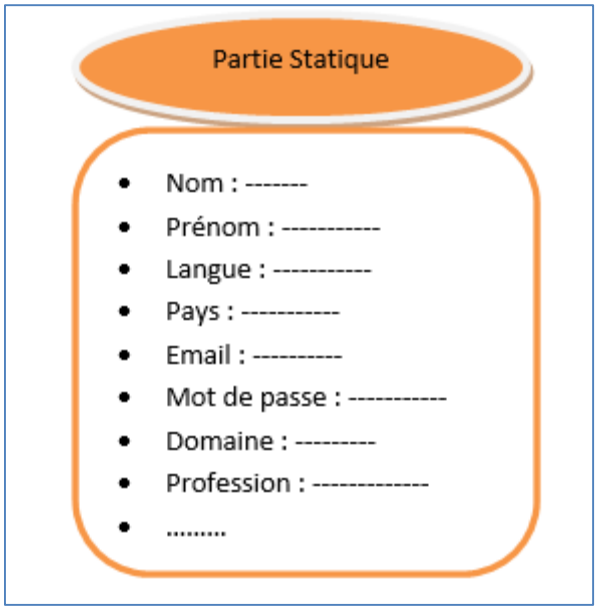


Figure 4.4: présentation de la partie statique

3.2 Partie dynamique :

Cette composante regroupe l'ensemble des informations collectées sur le comportement de l'utilisateur. Elle sert pour sauvegarder l'historique pertinent de l'utilisateur et elle comporte les documents jugés pertinents par l'utilisateur, plusieurs facteurs, y compris la fréquence des visites sur une page, le temps passé sur la page (longuement visités, trop cliqués, sauvegardés, imprimés, …). Ces documents sont utilisés pour extraire le centre d'intérêts de l'utilisateur (L'ensemble de termes les plus pondérés dans l'historique forme le centre d'intérêts).

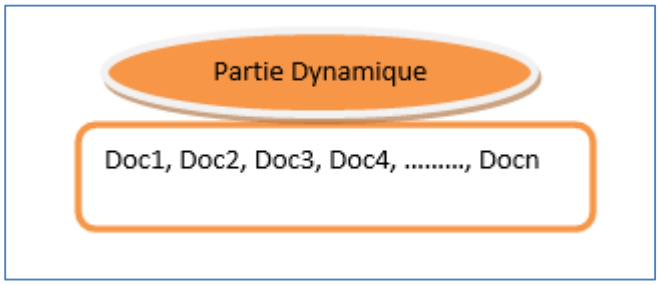
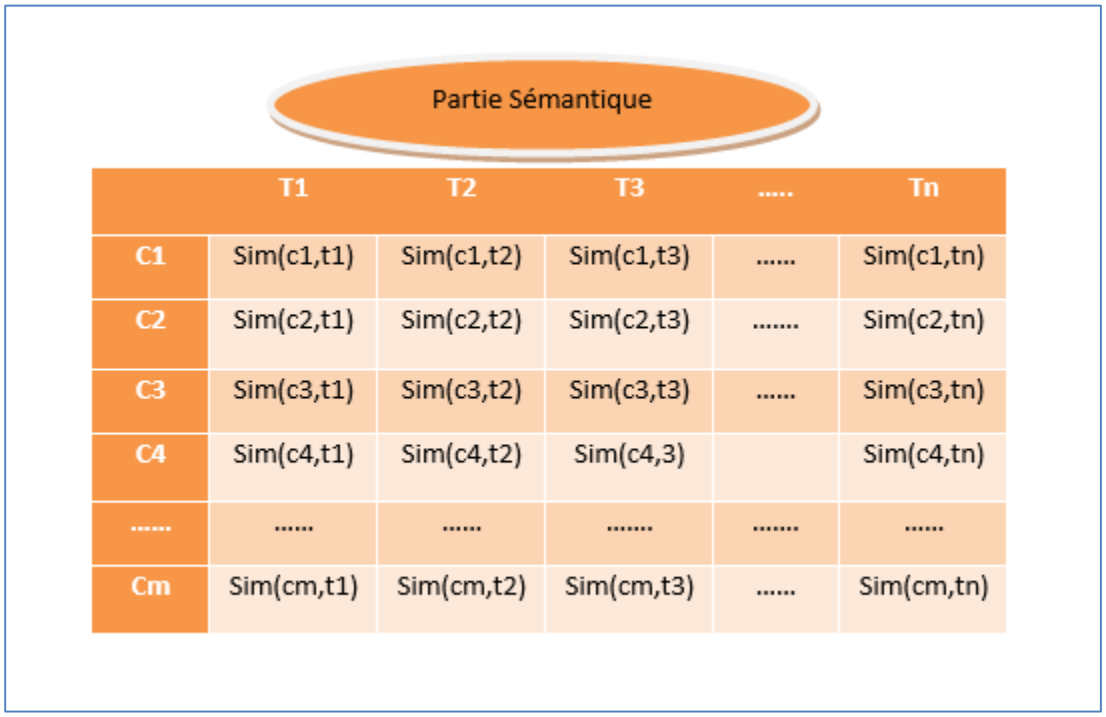


Figure 4.5: présentation de la partie dynamique

3.3 Partie sémantique :

Le profile utilisateur sémantique est représenté par un ensemble des termes extrait des documents pertinents de l'historique de recherche et qui peut être changé au cours des sessions de recherche, l'importance de chaque terme est définie par une pondération. Pour extraire les concepts utilisés pour la désambiguïsation sémantique de ces termes, on va les projetés sur une ontologie.



	T1	T2	T3	.....	Tn
C1	Sim(c1,t1)	Sim(c1,t2)	Sim(c1,t3)	.....	Sim(c1,tn)
C2	Sim(c2,t1)	Sim(c2,t2)	Sim(c2,t3)	.....	Sim(c2,tn)
C3	Sim(c3,t1)	Sim(c3,t2)	Sim(c3,t3)	.....	Sim(c3,tn)
C4	Sim(c4,t1)	Sim(c4,t2)	Sim(c4,3)		Sim(c4,tn)
.....	.....	.....	.....	.....	.....
Cm	Sim(cm,t1)	Sim(cm,t2)	Sim(cm,t3)	.....	Sim(cm,tn)

Figure 4.6: présentation de la partie sémantique

3.3.1 L'extraction et la pondération des termes de la partie dynamique

Notre approche consiste d'extraire tous les termes t_i descriptifs (mots simples) des documents pertinents de l'historique de l'utilisateur. Nous procédons à l'extraction des

termes en appliquant les outils de TAL traditionnellement utilisés en RI, ce qui correspond à :

- L'élimination des mots vides (stop lists), c'est à dire les listes prédéfinies qui contiennent les mots qui peuvent être ignorés.
- Lemmatisation ou radicalisation, prendre la forme canonique des mots (lemme).
- Une bonne approximation de l'importance du terme dans le document est donnéepar la combinaison des deux facteurs (TF * IDF), qui est la fonction de pondération laplus répondue.

$$w(t) = \frac{tf}{max\ tf} * \log(1 + \frac{N}{n}) \qquad (4.1)$$

Avec:

tf : fréquence totale de terme *t* dans les documents de l'historique.

Max tf : la plus grande fréquence des termes dans les documents de l'historique.

N : nombre de documents de l'historique.

n : nombre de documents contenant le terme *t* de l'historique.

3.3.2 La désambiguïsation des termes

Notre approche présente une évaluation des liens sémantique entre les termes pertinents de profile utilisateur et les concepts fournie par l'ontologie WordNet. L 'utilisation de la sémantique sert à identifier des concepts les plus proches pour désambiguïser les termes. L'ontologie wordnet peut fournir plusieurs concepts pour un terme.

Chaque terme t appartient au profile utilisateur sera projeté sur l'ontologie WordNet dans le but d'extraire les concepts proches du terme t, ensuit les concepts retenus seront pondéré par une mesure de similarité sémantique locale qui représente son importance locale par rapport au terme relativement à son contexte dans le document, et d'autre représentant son importance globale par rapport au terme relativement à son contexte dans tous les documents pertinents.

Parmi les concepts (synsets) correspondants au terme t_i dans WordNet, nous retenons seulement ceux ayant une similarité locale non nulle (SimLoc (t_i, c_j) > 0).

3.3.2.1 Pondération locale d'un concept.

Pour représenter l'importance locale (l'importance dans le document) d'un concept $c_j$retenu pour un terme t_i par rapport à son contexte locale, nous associons à chaque concept retenu (synset de WordNet) un score basé sur une mesure de similarité sémantique inspirée de la formule de lesk (M, 1986)
Cette formule compare les glossaires (définitions) d'un concept et le contexte d'un terme, et calcule un score en comptant le nombre de mots communs entre eux.

$$SimLoc(t_i, c_j) = \frac{|ctxt(t_i) \wedge \text{gloss}(c_j)|}{|ctxt(t_i) \vee \text{gloss}(c_j)|^2} \quad (4.2)$$

Avec :

- gloss (c_j) : l'ensemble des mots constituant la définition et les différentes relations sémantique de concept c_j (hyperonymes, hyponymes, …) dans *WordNet*.
- ctxt(t_i) : la partie de discours ou la phrase (part of speech) de terme t_i dans le document, noté :

$ctxt\ (\mathbf{t_i}) = m1, m_2, \ldots, t_i, \ldots . m_{j=|\ ctxt\ (\mathbf{ti})|}$

3.3.2.2 Pondération globale d'un concept.

L'importance globale d'un concept $c_j$retenu pour un terme pertinent $t_i$reflète son importance globale dans l'ensemble des documents pertinents de la collection (l'historique). Cette similarité est calculée par la formule suivante :

$$\text{SimGlob}\ (t_i, c_j) = \frac{\sum_{j=1\ldots|d(ti,cj)|} \text{SimLoc}\ (t_i, c_j) * |d(t_i, c_j)|}{|d(t_i)|^2 + |d(c_j) - d(t_i, c_j)|^2} \quad (4.3)$$

Avec :

- *Simloc* (t_i, c_j) : la similarité locale entre $t_i$et $c_j$dans le j^{eme} document.
- $d(t_i, c_j)$| : le nombre de documents dans lesquels $t_i$est désambigüisé par c_j.
- $|d(t_i)|$: le nombre de documents contenant t_i dans tout l'historique de recherche.
- $|d(c_j)|$ *:* le nombre de documents contenant des termes désambigüisé par c_j dans tout l'historique de recherche.
- $|d(c_j)-d(t_i, c_j)|$ *:* le nombre de documents contenant des termes désambiguïsé par c_j et ne contenant pas $t_i$dans tout l'historique de recherche.

4. Processus de recherche d'information personnalisée :

Notre processus de recherche sert a reformuler la requête envoyée par l'utilisateur. On prendre en compte les éléments de profile de l'utilisation dans la requête avant son exécution afin de proposer des résultats plus pertinents du point de vue de l'utilisateur, et pour une amélioration des performances du système.

Notre processus d'enrichissement de la requête à l'aide du profil utilisateur, permet de prendre en charge le centres d'intérêts de l'utilisateur et demieux expliquer ses besoins. L'enrichissement sert à ajouter des termes susceptibles afin d'améliorer la correspondance avec l'intention de l'utilisateur, principalement se fait en deux étapes :

1)- la recherche des termes ou concepts qui sont en relation avec la requête et qui ne sont pas contradictoires avec elle (qui maximise la similaritéentre les termes de la requête et de centre d'intérêts).

2)- intégration de ces termes à la requête.

Nous présentons d'abord une description fonctionnelle qui illustre notre processus d'accès personnalisé :

1. Le système reçoit la requête utilisateur Q.
2. Sélectionner parmi les concepts du centre d'intérêts les concepts les plus adéquate (similaire) à chaque terme de la requête.
3. Reformuler la requête Q en ajoutant lesconcepts sélectionnés C.
4. Envoi de la requête reformulée Q' au moteur de recherche.
5. Les résultats retournés par le moteur de recherche seront transmis à l'utilisateur.
6. Le système observe les réactions et les interactions de l'utilisateur avec les résultats mis à sa disposition, pour mettre à jour son profil (identifier les nouveaux termes et concepts pertinents).

4.1 L'enrichissement de la requête :

La requête initiale de l'utilisateur est reformulée dans l'objectif d'élargir le champ de la recherche pour sélectionner plus de résultats pertinents sans aller plus loin que les intérêts de l'utilisateur. La reformulation consiste à enrichir la requête initiale *Q* par l'injection d'un concept ou terme issu du centre d'intérêts.

Donc, la nouvelle requête reformulée Q' est obtenue sur la base de la requête initiale Q suivant les deux méthodes suivantes :

Notre approche propose deux scenario pour l'ajout des nouveauxconcepts a la requête initiale et faire une comparaison entre eux dans la phase d'expérimentation.

Scenario1 :

Le nombre de concepts à rajouter est égal au maximum au nombre de termes dans la requête. Dans ce cas on ajoute a chaque terme de la requête le concept le plus similaire

Unerequête Q initial qui contient des termes {t1,t2,t3,….,tm}, chaque terme ti présente dans le profil utilisateur a des concept similaires extrait de l'ontologie WordNet {c1,c2,…,cm}, lors de la processus de reformulation, le système ajoute a chaque terme le concept adéquat qui maximise la similarité sémantique globale (4.3) .

La requête finale serait comme suite :

Q'= {t1,c1,t2,c2,……,tm,cm}

Les termes de la requête initial qui ne se trouve pas dans le profil utilisateur n'ont pas des concepts a ajouté a la requête finale.

$$Q'=Q+Ci$$

Avec :

Q' : la nouvelle requête reformulée.
Q : la requête initiale.
Ci : ensemble des concepts les plus similaires aux termes de la requete {c1,c2,……cn}

Scenario 2 :

Le concept a ajouté est celui qui maximise la somme des similarités entre le conceptet tous les termes de la requête initial.

Une requête Q initial qui contient des termes {t1,t2,t3,….,tm}, chaque terme ti présente dans le profil utilisateur a des concept similaires extrait de l'ontologie WordNet {c1,c2,…,cm},lors de la processus de reformulation, le système ajoute un seul concept $\text{Best}(c_j)$.

$$\text{Sim}(Q, c_j) = \sum_{t \epsilon Q} \text{SimGlob}\ (t, c_j) \quad (4.4)$$

$$Q'= Q + C$$

Avec :

Q' : la nouvelle requête reformulée.
Q : la requête initiale.
C : le concept qui maximise la relation de similarité avec les termes de la requête Q selon la formule (4.4).

4.2 Mise à jour de profil :

Un profil utilisateur ne peut pas être considéré stable dans le temps, une mise à jour de profil est donc requise. La mise à jour du notre profil sémantique doit assurer d'un côté, son enrichissement et son évolution dans le temps par l'addition des nouveaux termes pertinents, leurs poids ainsi que leurs annotations sémantique (synsets annoté) dans la représentation du profil utilisateur après un traitement prédéfini. Et d'un autre côté, elle doit veiller sur l'optimisation de l'espace occupé et le temps de réponse pris par les différents calculs, et ce par l'élimination des termes jugés à un certain moment non pertinents, et des documents dont les termes n'apparaissent pas dans le centre d'intérêts pendant une temps définie.

4.2.1 L'alimentation de profil (l'ajout).

Après mettre les résultats finaux à la disposition de l'utilisateur, notre système observe l'interaction de l'utilisateur avec ces résultats pour identifier les nouveaux documents jugés, explicitement ou implicitement, pertinents. Ces documents sont indexés, leurs termes sont extraits et pondérés. Ensuite l'ensemble de termes sont projeté sur l'ontologie (WordNet), et aussi les concepts utilisées pour désambiguïser les termes sont filtrés selon son domaine d'usage et pondérés selon la formule (4.2) et (4.3). Ces termes et concepts sont enregistrés dans le centre d'intérêts.

4.2.2 Epuration de profil (la suppression) :

Pour filtrer et alléger l'historique de recherche de l'utilisateur, notre stratégie consiste à supprimer tout document de l'historique de recherche ayant plus d'un an de ses indexations, Et ainsi les termes extraits à partir d'un document supprimé sont supprimés aussi. Et par transitivité, les concepts, associés seulement à un terme supprimé seront supprimés aussi.

5. Evaluation de l'approche :

Comme nous avons vus dans le chapitre 3 (la recherche d'information personnalisée), l'évaluation d'une approche de RIP est une tâche très problématique. En raison du caractère subjectifs des utilisateurs, de fait qu'elle dépend en particulier de l'utilisateur qui est le seule qui peut juger la pertinence des résultats de recherche selon son point de vue. Vu qu'aucune collection de test standard n'a été construite à notre connaissance pour évaluer l'efficacité de l'accès personnalisé à l'information, les auteurs des approches de RIP ont recours à développer leurs propres solutions, ce qui rend la tâche de comparaison des performances problématique aussi.

D'après notre point de vue, l'exploitation des collections de tests avec des requêtes et des pré-jugements de pertinence effectués par des experts, comme c'est fait dans la majorité d'approches de la RIP, ne résout pas le problème car il ne s'agit pas de l'avis de l'utilisateur en question.

Nous avons choisi d'évaluer notre approche à l'aide d'une simulation. Dans cette simulation nous avons donc suivi la stratégie suivante.

5.1 Stratégie d'évaluation :

nous conduirons notre phase d'évaluation selon les étapes suivantes:

- *Construction du profil de l'utilisateur* :nous avons construit différemment 5 profils pour 5 utilisateurs de différents domaines professionnels ayant des intérêts différents. Nous avons construit chaque profil avec la collaboration de l'utilisateur concerné.
- *Requêtes et recherche* : Nous utilisons *Google* comme moteur de recherche. Chaque utilisateur a envoyé 5 requêtes, parmi ces requêtes au moins une requête est formulée de façon à exprimer un besoin informationnel spécifique. Certaines requêtes des utilisateurs sont communes.

- *Evaluation*: pour obtenir un réel feedback, les jugements de pertinence concernant les résultats renvoyés sont effectués manuellement par les utilisateurs. L'évaluation consiste à calculer le facteur de X-précision (taux de documents pertinents retournés) aux 5, 10 et 15 documents.

- *Comparaison des performances*: pour comparer la contribution de notre approche(les deux scénarios proposé) par rapport aux approches classiques, , nous calculons la précision selon les 3 scénarios suivants :

  (1) requête envoyée sans reformulation,
  (2)requête reformulée, en utilisant notre profil sémantique (scénario 1)
  (3) requête reformulée, en utilisant notre profil sémantique (scénario 2)

5.2 L'expérimentation :

L'expérimentation consiste à exécuter les requêtes des utilisateurs sur notre modèle de recherche selon les trois scénarios: (1) requête sans reformulation, (2) et (3) requête reformulée à l'aide de notre profil sémantique en utilisant deux scénarios. L'objectif est de démontrer que notre modèle de recherche personnalisé sémantique basé sur la reformulation de requête à partir d'historique de recherche, est plus performant que les modèles de recherche classique basé sur les mots clés. et de comparer la performance des deux scénario proposés.

5.2.1 Implémentations

5.2.1.1 Construction des profils et requêtes.

La première étape de notre expérimentation est la construction des profils utilisateurs qui définissent les centres d'intérêts de l'utilisateur, ainsi que les requêtes à envoyées pour répondre aux leurs besoins informationnels. Nous avons choisi d'exprimer ses données en anglais. Tableau suivant présente le profil de chaque utilisateur :

Utilisateur	Requêtes	Utilisateur	Requêtes
User 1 Nom : amirat Prénom :sarah Age :24 ans Domaine :informatique	**Q 1:**machine learning. **Q2** :Swift. **Q3** :programming language. **Q4** : narrow AI. **Q5** :Ruby.	User 3 Nom : boudekkas Prénom : bouchra Age :24 ans Domaine :fashion	**Q1** : TRENDS. **Q2** : suit. **Q3**: feminism looks. **Q4 :** style **.** **Q5** : classic pieces.
User 2 Nom : cheriet Prénom : asrar Age :23 ans Domaine :informatique	**Q1 :**database **Q2 :**python **Q3:** java compiler **Q4 :**database.management **Q5 :** python code.	User 4 Nom : cheriet Prénom : lotfi Age :24 ans Domaine : games	**Q1**:Subway. **Q2**: Player Unknown. **Q3**:Garena. **Q4**: battle royale **.** **Q5**: PUBG lite.
User 5 Nom : amrani Prénom : amel Age :24 ans Domaine :cuisine	**Q1**: fool **Q2**: key lime **Q3**:traditional plate **Q4**:meet burger **Q5**:healthy food		

Table 4.1 : Les requêtes de test de chaque utilisateur

User	**Termes**	**Concept (sim)**
Nom : boudekkas **Prénom : bouchra** **Age : 23 ans**	Designer	Interior_designer (0.0)
		Graphic_designer (0.0019)
		Couturier (0.0057)
	boots	Boot (0.0038)
		Bang (0.0)

Domaine : fashion		Kick (0.0)
	Dresses	Dresse (0.0057)
		Attire (0.0015)
		Apparel (0.0038)
		Preen (0.0038)
		Snip (0.0018)
		Trim (0.0)
	Pull	Wrench (0.0)
		Puff (0.0)
		Attract (0.0)
		Draw (0.0057)
		Extract (0.0076)
	Style	Espressive_style (0.0095)
		Vogue (0.0076)
		Dash (0.0019)
		Stylus (0.0)
	Trends	Trendency (0.0019)
		Course (0.0)
		Vogue (0.0076)
		Swerve (0.0019)

Table 4.2 : exemple sur Les données des profils utilisateurs.

5.2.1.2 L'environnement de développement :

Chaque projet d'informatique dépend sur plusieurs ressources et outils car chaque ressource va nous aider dans un point donnée, pour cela nous présentons dans cette section les différents langages de programmation ou outils utilisé dans notre approche :

- Python :

Python est un langage de programmation open source **créé par le programmeur Guido van Rossum en 1991**. Il tire son nom de l'émission Monty Python's Flying Circus.
Il s'agit d'un **langage interprété**, qui ne nécessite donc pas d'être compilé pour fonctionner. Un programme » interpréteur » permet d'exécuter le code Python sur n'importe quel ordinateur. Ceci permet de voir rapidement les résultats d'un changement dans le code. En revanche, ceci rend ce langage plus lent qu'un langage compilé comme le C.
En tant que l**angage de programmation de haut niveau**, Python permet aux programmeurs de se focaliser sur ce qu'ils font plutôt que sur la façon dont ils le font. Ainsi, écrire des programmes prend moins de temps que dans un autre langage. Il s'agit d'un langage idéal pour les débutants.https://www.lebigdata.fr/python-langage-definition
Le langage Python doit sa popularité à plusieurs avantages qui profitent aussi bien aux débutants qu'aux experts. Tout d'abord, il est **facile à apprendre et à utiliser**. Ses caractéristiques sont peu nombreuses, ce qui permet de créer des programmes rapidement et avec peu d'efforts. De plus, sa syntaxe est conçue pour être lisible et directe.
Un autre avantage du Python est sa popularité. Ce langage **fonctionne sur tous les principaux systèmes d'exploitation** et plateformes informatiques. De plus, même s'il ne s'agit clairement pas du langage le plus rapide, il compense sa lenteur par sa versatilité.
Enfin, même s'il est principalement utilisé pour le Scripting et l'automatisation, ce langage est aussi utilisé pour **créer des logiciels de qualité professionnelle**. Qu'il s'agisse d'applications ou de services Web, le Python est utilisé par un grand nombre de développeurs pour créer des logiciels.(webresource)

5.3 Résultats et discussion :

Users	p@5			p@10			p@15		
	Rnr	s1	s2	Rnr	s1	s2	Rnr	s1	s2
User1 Nom : amirat Prénom : Sarah Domaine : informatique	0.2	1	1	0.2	1	1	0.2	1	1
	0.6	0.8	0.8	0.7	0.9	1	0.7	0.8	0.9
	0.0	1	1	0.0	1	1	0.0	1	1
	0.4	1	1	0.5	1	0.9	0.46	0.73	0.86
	0.4	0.8	0.6	0.7	0.8	0.7	0.73	0.73	0.73
X-précision moyenne	0.32	0.92	0.88	0.42	0.94	0.92	0.42	0.85	0.89
User2 Nom : cheriet Prénom : asrar Domaine : informatique	0.0	0.8	0.8	0.0	0.9	0.9	0.0	0.8	0.9
	0.2	0.6	0.6	0.3	0.6	0.9	0.39	0.6	0.6
	0.4	0.6	0.4	0.5	0.8	0.4	0.53	0.73	0.26
	0.2	0.6	0.6	0.5	0.3	0.6	0.4	0.33	0.73
	0.4	0.2	0.8	0.2	0.2	0.9	0.33	0.2	0.8
X-précision moyenne	0.24	0.56	0.64	0.3	0.44	0.74	0.33	0.53	0.66
User3 Nom : amrani Prénom : amel Domaine : cuisine	0.0	1	1	0.0	1	1	0.0	0.87	0.87
	0.4	1	0.8	0.6	1	0.9	0.6	1	0.93
	0.0	0.0	0.6	0.0	0.0	0.7	0.0	0.0	0.8
	0.2	1	1	0.2	1	1	0.33	1	1
	0.2	0.9	1	0.3	0.9	1	0.33	0.8	1
X-précision moyenne	0.16	0.78	0.88	0.22	0.78	0.92	0.25	0.73	0.92
User4 Nom : boudekkas Prenom : bouchra Domaine : fashion	0.0	1	1	0.0	1	1	0.0	1	1
	0.0	0.8	0.8	0.1	0.8	0.8	0.13	0.86	0.86
	0.2	0.6	1	0.3	0.8	1	0.33	0.85	1
	0.0	1	1	0.1	1	1	0.13	1	1
	0.0	0.2	0.8	0.0	0.3	0.7	0.66	0.33	0.6
X-precision moyenne	0.04	0.72	0.92	0.1	0.78	0.9	0.25	0.8	0.89
User5 Nom : cheriet Prénom : lotfi Domaine : games	0.0	1	1	0.0	1	1	0.13	0.93	0.93
	0.0	0.8	0.8	0.1	0.9	0.9	0.2	0.93	0.93
	0.0	1	0.6	0.0	1	0.6	0.0	1	0.53
	0.6	0.8	0.6	0.6	0.7	0.8	0.6	0.67	0.8
	0.8	0.4	1	0.5	0.4	0.9	0.6	0.33	0.8
X-précision moyenne	0.28	0.8	0.8	0.24	0.6	0.84	0.31	0.77	0.79

X-précision moyenne générale	0.21	0.76	0.82	2.26	0.7	0.86	0.31	0.74	0.83

Table 1 **:** Résultats expérimentaux des requêtes avant et après la reformulation pour chaque utilisateur en utilisant les deux scénarios

La première observation au vu de ces résultats est que notre modèle (que ca soit le scénario 1 ou 2) est efficace et présente des améliorations de performances significatives comparativement au modèle classique (requête sans reformulation) aux points de précision p@5, p@10 et p@15.

On a observé que le degré d'amélioration de précision se change d'une requête a une autre et cela dépend de la correspondance entre les centres d'intérêts de l'utilisateur simulés et la requête (pour les deux scénarios de reformulation).

On vient de comparer les deux scénarios de reformulation sémantique personnalisé , on observe que les deux scénarios améliore les résultats obtenu au utilisateur ,ms le scénario 2(l'ajout de concept qui maximise la similarité avec les termes de la requête)est plus efficace par rapport au scénario 1(l'ajout de tous les concepts des termes de requête) , et cela est dû au fait que certain termes de la requête n'exprime pas exactement ce que l'utilisateur souhaite et donc lors de l'ajout de concepts de ces termes la précision de recherche se démunie

Voici quelques exemples d'améliorations des résultats obtenus après la reformulation de requête apportés dans notre approche de personnalisation sémantique :

- **Exemple 1** : Les premiers résultats de la requête « fool » envoyé par l'utilisateur 1 qui s'intéresse au informatique ne satisfaire pas ses besoins.

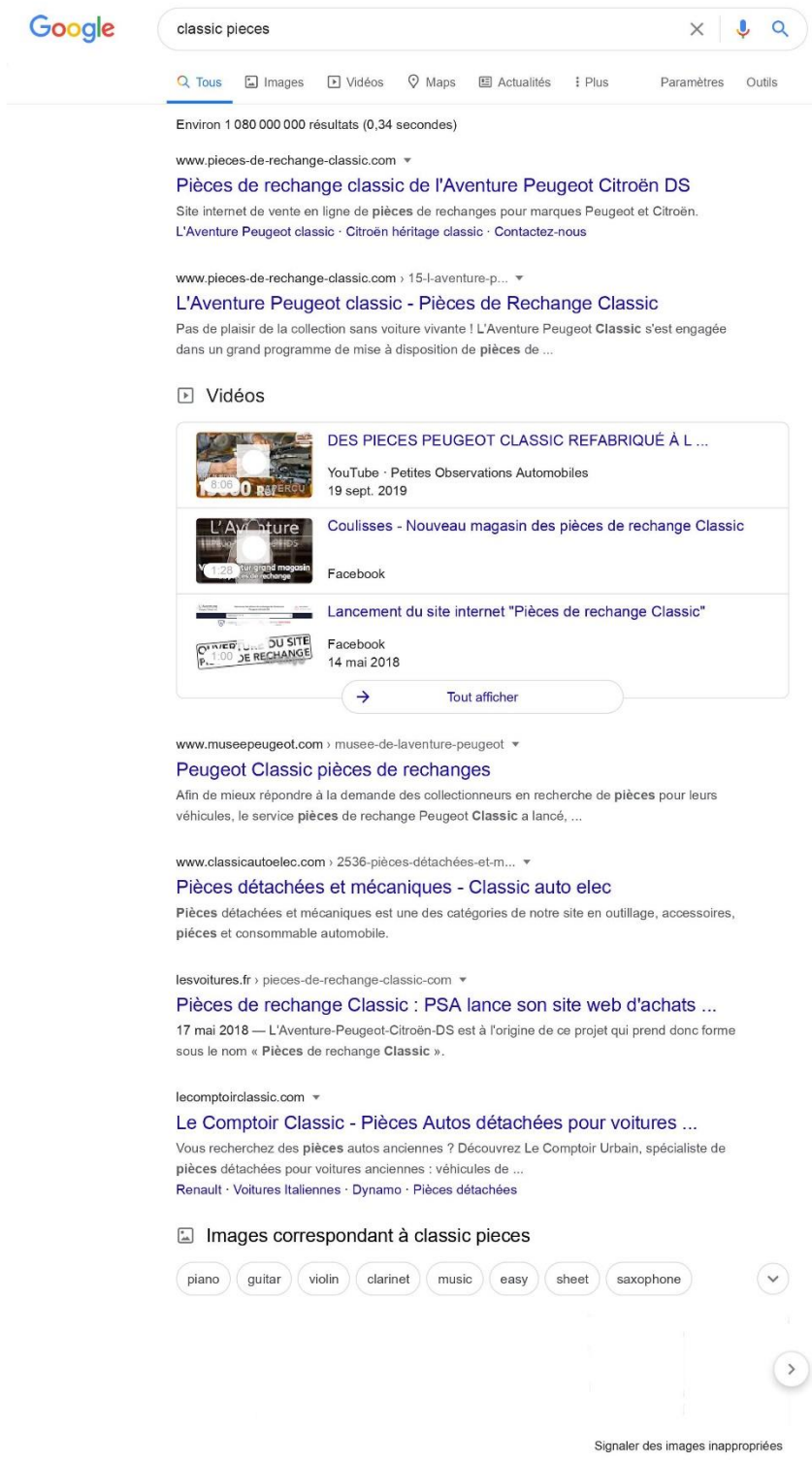


Figure 4.7: représentation des résultats de la requête « pièce classique » avant la reformulation.

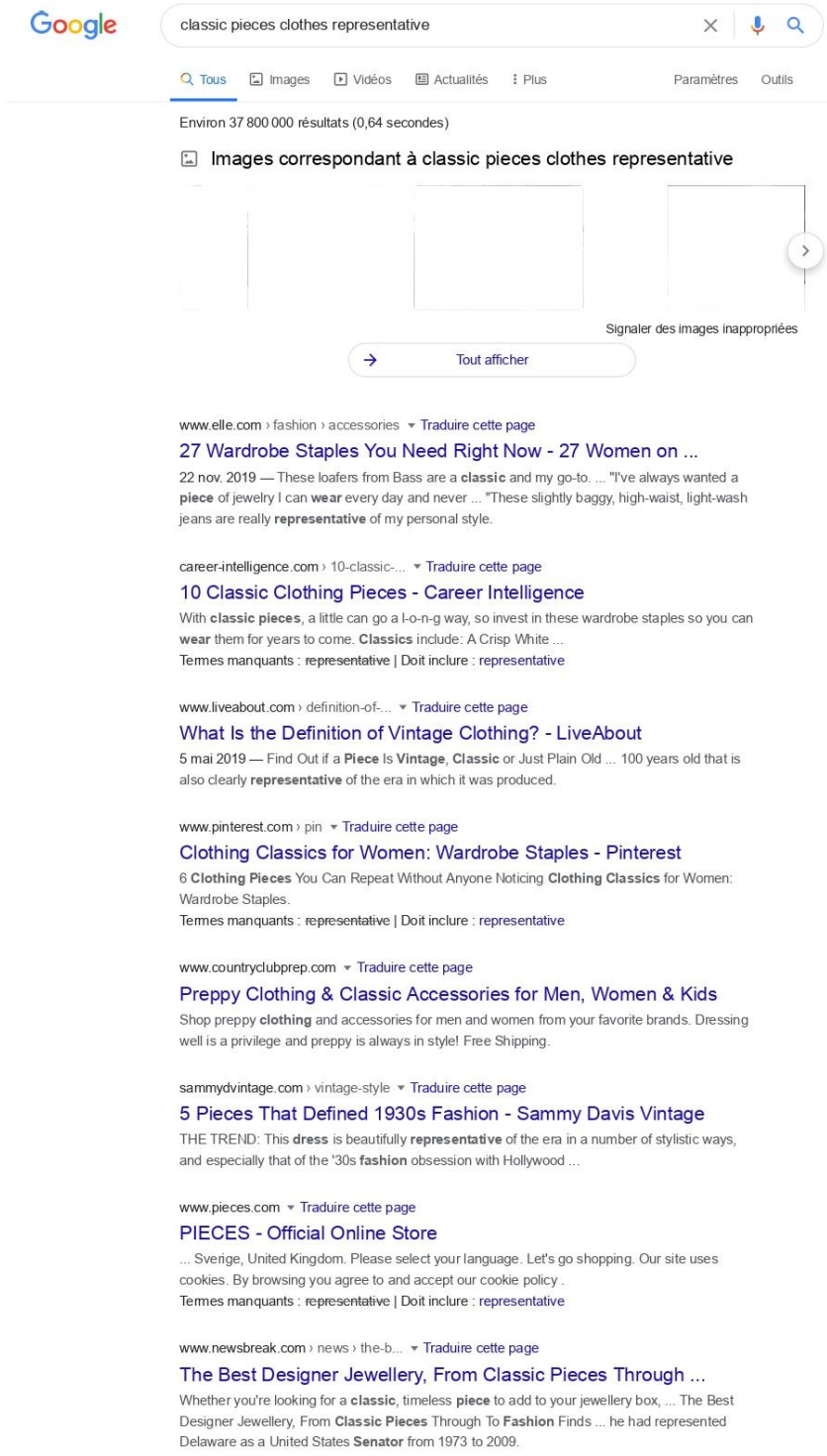


Figure 4.8:représentation des résultats de la requête « piece classique » après la reformulation en utilisant le 1er scénario.

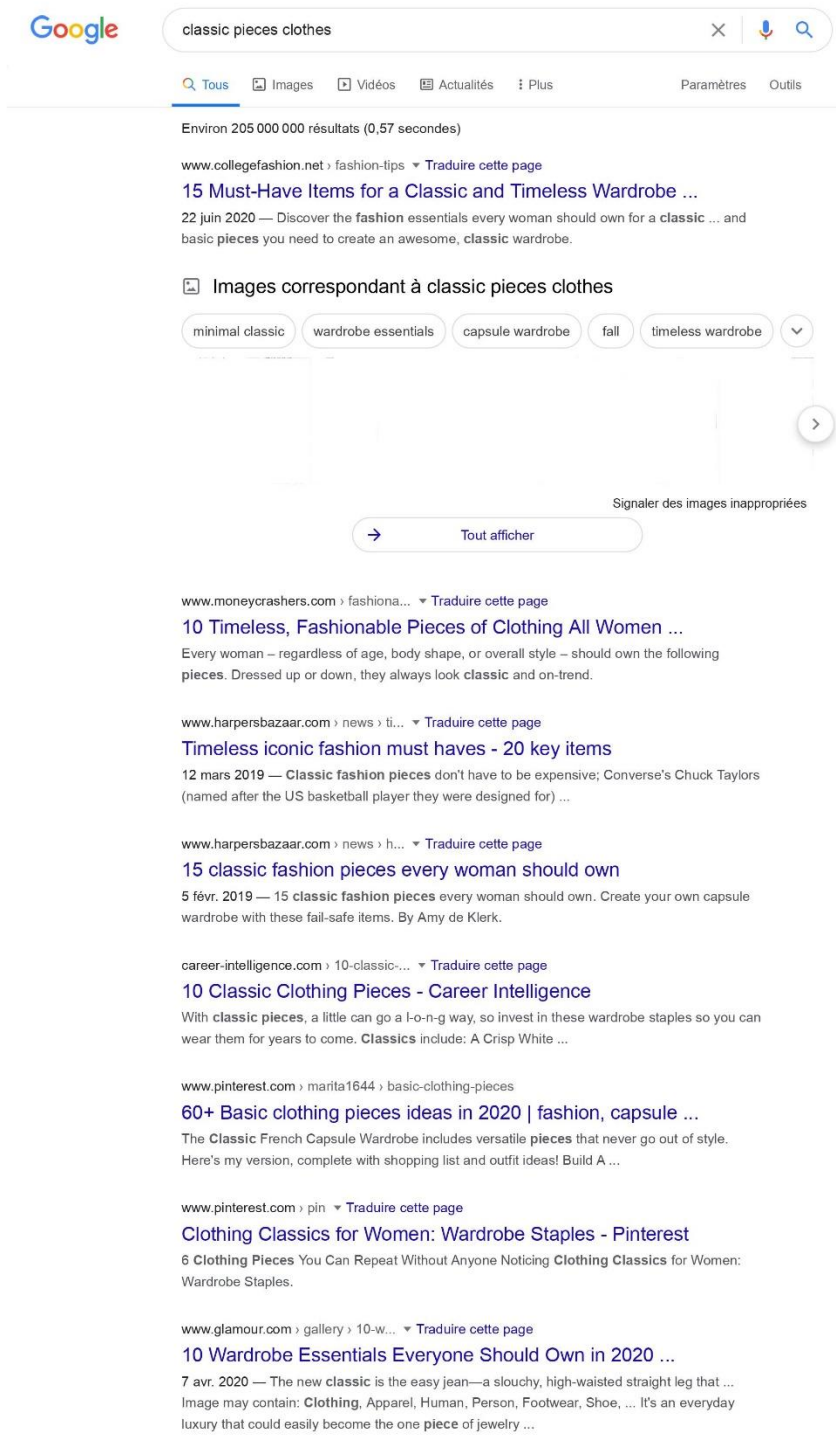


Figure 4.9 : représentation des résultats de la requête « pièce classique » après la reformulation en utilisant le 2ème scénario.

6. CONCLUSION :

Dans ce chapitre nous avons présenté notre approche qui est basée sur un profil d'utilisateur sémantique qui s'intéresse aux l'historique de l'utilisateur ainsi à une ontologie sémantique, pour réaliser un système de recherche d'information d'accès personnalisé. Aussi nous avons présentées dans ce travail désambiguïsation des termes de l'historique pertinent de recherche sur l'ontologie dans le but d'extraire les concepts les plus similaire aux besoin d'utilisateur.

Cette approche est proposée dans l'objectif d'améliorer la pertinence des résultats de recherche retournés, par la reformulation de la requête initiale par l'intérêt de l'utilisateur.

A la fin nous avons présentées une méthode d'évaluation des performances de notre système SRIPS. Les résultats obtenus montrent que plus que le profil est riche, plus la requête est mieux reformulée et plus les résultats sont pertinents.

Conclusion générale

Notre travail se situe dans le contexte de la recherche d'information, le domaine de la recherche d'information a élargi ces dernières années en direction d'application ,visant non seulement à retrouver des éléments à partir d'une requête initiale de l'utilisateur mais à chercher de l'information répondant au mieux à la demande de l'utilisateur ,donc la plus adaptée à ses besoins, la solution consiste à intégrer le profile utilisateur dans le processus de recherche(recherche d'information personnalisée)

Le but de notre travail est de proposer une solution pour personnaliser le processus de recherche d'information en prenant en considération le profil utilisateur qui contient son historique de recherche

Afin d'arriver à cet objectif nous avons au cours de ce travail passé en revue de notions théorique entourant le sujet ,nous avons commencé par les notions de la recherche d'information classique et l'utilisation de la sémantique dans les systèmes de recherche d'information classique , puis les notions principaux de la recherche d'information personnalisée ,la modélisation du profil utilisateur et son intégration au sein d'un système de recherche d'information et l'utilisation de la sémantique dans les systèmes de recherche d'information personnalisée. Nous avons terminé par la reformulation de la requête dans le système de recherche d'information classique, sémantique et personnalisé

Au terme de notre travail nous avons intégré le profil utilisateur au sein de système de recherche d'information en prenant en compte les termes qui caractérisent les documents de l'historique de recherche et les concepts extrait d'une ontologie sémantique pour désambigüise les termes. Ces termes et concepts permettent de renvois plus de documents pertinentes et donc mieux rependre au utilisateur.

Dans un deuxième temps nous avons décrit les modules qui composent notre système RIPS, à savoir le module d'acquisition (construction) du profil utilisateur qui a pour but d'instancier sa représentation définie au créable, le module de recherche qui prend en compte la requête émet par l'utilisateur et son profil sémantique afin de retourner les éléments qui répondent aux besoins de l'utilisateur et enfin le module de feedback qui permet de mettre à jour le profil utilisateur

Des simulations ont été effectués afin d'évaluer notre system de recherche d'information personnalisée sémantique. Les résultats obtenus sont satisfaites nous croyons bien que nous avons bien atteint les objectifs que nous nous sommes fixés au départ.

REFERENCES :

[Ameni YENGUI, 2016]

Ameni YENGUI, Système de recherche d'information sémantique pour les bases de visioconférences médicales à travers les graphes conceptuels, juin 2016.

[Avigli and Velardi2003]

Roberto Navigli and Paola Velardi. An analysis of ontology-based query expansion

Strategies , 2003.

[Bouramoul, 2011]

MR. ABDELKRIM BOURAMOUL**,** Recherche d'information Contextuelle et Sémantique sur le web, 2011.

[Bouhini, 2014]

Chahrazed Bouhini, Impact des réseaux sociaux sur le processus de recherche d'information, 2014.

[B.Tan, X.Shen, and C.Zhai, 2006]

B.Tan, X.Shen, and C.Zhai. Mining long-term search history to improve search

Accuracy. In Proceedings of the 12th international conference on Knowledge Discovery and Data Mining, University of Illinois at Urbana-Champaign, pages 718–723, August2006.

[Bradley, 2000]

Case-Based User Profiling for Content Personalization, K. Bradley, R Rafter, B. Smyth, Lecture Notes in Computer Science, 2000.

[Bissan Audeh, 2014]

Bissan Audeh, Reformulation sémantique des requêtes pour la recherche d'information ad hoc sur le Web, 2014

[C.VAN Rijsbergen, 1986]

C. Van Rijsbergen. A non-classical logic for information retrieval. *The computer*

Journal, 29(6):481.485, 1986.

[Chi, 2001]

Stuart K. Card, Peter Pirolli, Mija Van Der Wege, Julie B. Morrison, Robert W. Reeder, Pamela K. Schraedley, Jenea Boshart, Information Scent as a Driver of Web Behavior Graphs: Results of a Protocol Analysis Method for Web Usability,2001.

[Cronen-Townsend et al, 2002].

Cronen-Townsend. Predicting query performance. Research and development in information retrieval, 2002.

[Cui et al, 2003]

Hang Cui, Ji-rongWen, Jian-yun Nie, andWei-ying Ma. Query expansion by mining

User logs, 2003.

[Caro, 97]

Budi Yuwono, Savio L.Y. Lam, Jerry H. Ying, Dik L. Lee, A World Wide Web Ressource Discovery System, 1997.

[Daoud ,2009]

Mariam DAOUD, Accès personnalisée a l'information : approche basée sur l'utilisation d'un profil utilisateur sémantique dérivé d'une ontologie de domaines `a travers l'historique des sessions de recherche,2009

[Farzan *et al.*, 2006],

Farzan R., Brusilovsky P., « Social Navigation Support in a Course Recommendation System », Proc. AH 2006, LNCS 4018, Springer, 2006

[H. Aliane, 2004]

H. Aliane, Un Système de reformulation de requêtes pour la recherche d'information, 2004.

[Hammech, 2013]

M. Hammache Arezki, Recherche d'Information : un modèle delangue combinant mots simples et mots composés ,2013.

[Haase *et al.*, 2005]

Haase, P., Hotho, A., Schmidt-Thieme, L., Sure, Y., « Collaborative and Usage-driven evolution of personalized ontologies », in Proceedings of the 2nd European Semantic Web Conference, Heraklion, Greece, May 29-June 1 (2005) 486-499

[Ilhem Boussaid, 2005]

Ilhem Boussaid, Personnalisation de l'information et gestion des profils utilisateurs : une approche fondée sur les ontologies, 2005.

[Ingwersen, 1992]

Peter Ingwersen, From Information Retrieval to Information Interaction, January1992.

[Jean-Christophe BOTTRAUD, 2004]

Jean-Christophe BOTTRAUD, Un assistant adaptatif pour la recherche d'information : AIRA (Adaptive Information Retrievl Assistant), 2004.

[Kompaoré, 2008].

N.D.Y. Kompaoré. "Fusion de systèmes et analyse des caractéristiques linguistiques des requêtes: vers un processus de RI adaptatif». Thèse de doctorat en informatique, Université Paul Sabatier de Toulouse, 2008

[kechid, 2009]

Samir Kechid, Intégration du modèle utilisateur dans un système de recherche d'information distribuée, 2009.

[Krovetz et Croft, 1992]

Robert Krovetz, Croft, Lexical Ambiguity and Information Retrieval, 1992.

[Karine abbas2008]

Karine Abbas, Système d'accés personnalisé a l'information : application au Domaine médical, 2008.

[Lesk, 86]

M.E. Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from a nice cream cone. In Proceedings of the SIGDOC Conference, Toronto, 1986.

[Lechani et Boughanem, 2005]

Lechani et M. Boughanem, Accées personnalisé a l'information : vers un modèle basé sur les diagrammes d'influence ,2005.

[Lynda Tamine, 2000]

Lynda Tamine, OPTIMISATION DE REQUETES DANS UN SYSTEME DE RECHERCHE D'INFORMATION, 2000.

[Liu et al. 2004]

Liu et al , An effective approach to document retrieval via utilizing wordnet and recognizing phrases,2004.

[M. Boughanem and J. Savoy, 2008]

M. Boughanem and J. Savoy, editors. Recherche d'information ´etats des lieux et perspectives. Hermès Science Publications, http ://www.editionshermes.fr/, 2008.

[Maisonnasse, 2008].

Loïc Maisonnasse, Les supports de vocabulaires pour les systèmes de recherche d'information orientés précision : application aux graphes pour la recherche d'information médicale, 2008.

[Mobasher, 2000]

Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization, B. Mobasher, H. Dai, T. Luo, M. Nakagawa, Data Mining and Knowledge Discovery, 6(1):61-82, 2002.

[Middleton *et al.*, 2004]

Middleton, S.E., Shadbolt, N.R., De Roure, D.C., « Ontological User profiling in Recommender Systems », ACM Transactions on Information Systems, 22(1), 2004.

[MyriamHadjouni Krirn, 2012]

Myriam Hadjouni Krirn, Un système de recherche d'information personnalisée basé sur la modélisation multidimensionnelle de l'utilisateur, 2012.

[NesrinZemirli, 2008]

WAHIBA NESRINE ZEMIRLI, Modèle d'accès personnalisé à l'information basé sur les Diagrammes d'Influence intégrant un profil utilisateur évolutif,2008.

[Ourdia, 2011]

Mme 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,2011.

[Rami HARRATHI, 2010]

Rami HARRATHI, Recherche d'information conceptuelle dans les documents semi-structurés, 2010.

[Rocchio, 1971]

ROCCHIO J., Relevance feedback in information retrieval, Prentice Hall Inc., englewood cliffs, njédition, 1971

[Trichet *et al.*, 2003]

Trichet, F., Leclere, M., « A Framework for building competency-based systems », in Zhong N., et al. (Eds.), Proc. of ISMIS 2003, LNAI 2871, pp. 633–639, Springer, 2003.

[Tambellini, 2007].

C. Tambellini. "Un système de recherche d'information adapté aux données

Incertaines: adaptation du modèle de langue". Thèse de doctorat en informatique,

Université de Nice-Sophia Antipolis-UFR sciences, 2007.

[V.Raghavan and H.Sever, 1995]

V.Raghavan and H.Sever. On the reuse of past optimal queries. In Proceedings of

The 18th annual international ACM SIGIR conference on Research and development in information retrieval, NY, USA, pages 344–350, 1995

[Voorhees, 1993]

Ellen M. Voorhees, UsingWordNetTM to Disambiguate Word Senses for Text Retrieval, 1993.

[Wang and Zhai 2008]

Xuanhui Wang and ChengXiangZhai. Mining term association patterns from search

logs for effective query reformulation, 2008.

[Wilks et al., 90]

Y. Wilks, D. Fass, C. Guo, J.E. Mcdonald, T. Plate, B.M. Slator. Providing Machine Tractable Dictionary Tools. In Machine Translation, Vol.5, p.99-154. 1990.

[Zargayouna, 2005]

Haïfa Zargayouna, CatherineRoussey, Jean-Pierre Chevallet, Recherche d'information sémantique : état des lieux, 2005.

[Ziani, 2006]

Ziani, an Adaptation Approach: Query Enrichment by User Profile, 2006.

[warisali2.]
Django introduction Retrieved 20 october 2020, from https://developer.mozilla.org/en-US/docs/Learn/Server-side/Django/Introduction

[webresource]

python language definition Retrieved 23october , 2020, from https://www.lebigdata.fr/python-langage-definition